Listening. Learning. Leading.®

Research Memorandum
ETS RM–12-12

Statistical Report of Fall 2009 *CBAL*™ Reading Tests

Jianbin Fu

Maxwell Wise

Seunghee Chung

May 2012

Statistical Report of Fall 2009 *CBAL*™ Reading Tests

Jianbin Fu, Maxwell Wise, and Seunghee Chung
ETS, Princeton, New Jersey

May 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: James Carlson

Technical Reviewers: Peter van Rijn and Frank Rijmen

.

Abstract

In the Cognitively Based Assessment *of, for, and as* Learning (*CBAL*™) research initiative, we develop innovative K-12 prototype tests based on cognitive competency models. In this report, we present the statistical results of 2 CBAL Grade 7 reading tests administered to students in 9 states in fall 2009. Specifically, classical item statistics including rater reliability for human scored items, item P+, item-total correlations, item missing rates, differential item functioning, interscore correlations, and reliabilities of subscores and total scores were reported. Using item response theory, tests were calibrated and scaled based on the generalized partial credit model. Results from the concurrent calibration and the separate calibration with anchor linking were very similar and interchangeable. In addition, *t*-tests, multiple comparisons, and mixed models were used to examine the factors influencing test scores including demographic group, test form, test-order, student, and school. The results show that these 2 tests, as well as the linking sets, performed reasonably well.

Key words: CBAL, reading test, item analysis, item response theory, statistical report

Acknowledgments

The authors would like to thank Sharon Slater, Rick Morgan, Randy Bennett, Jim Carlson, Frank Rijmen, and Peter van Rijn for their helpful suggestions and edits on early versions of this report. Thanks are also due to Kim Fryer for her editorial assistance.

Table of Contents

List of Tables

Page

List of Figures

Page

The goal of the Cognitively Based Assessment of, for, and as Learning (*CBAL*™) research initiative is to create a model for an innovative K-12 assessment system that documents students' achievement (of learning), provides timely feedback information for education intervention (for learning), and is a worthwhile educational experience in and of itself (as learning) (Bennett, 2010). To help achieve these goals, CBAL summative tests are intended to be administered throughout a school year, with results aggregated for accountability purposes. This design is consistent with that of the Partnership for Assessment of Readiness for College and Careers, one of the two consortia funded by the federal government's Race to the Top assessment program (Robelen, 2010).

CBAL tests are developed based on underlying cognitive competency models that incorporate curriculum standards with results of learning sciences' research. The competency models describe skills that students need to learn and their interrelationships, for example, learning progressions (Deane, Fowles, & Bennett, 2009; O'Reilly & Sheehan, 2009a, 2009b). Tests are administered online and include innovative technology-enhanced items. These items are typically tied together by a common scenario that is intended to help measure higher-order, critical-thinking abilities.

This report presents the statistical results from administering two reading periodic accountability assessments (PAAs) to Grade 7 students in multiple states. In the following sections, the test and sampling designs are first introduced. Then the results from classical item analyses are reported, including rater reliabilities for human scored items, item P+ values, item-total correlations, item missing response rates, and differential item functioning. This description is followed by summary statistics of task scores, subscores and total raw scores including mean, standard deviation, interscore correlations, and reliabilities. Following that, item calibration and scaling results based on item response theory are reported. Results from both concurrent calibration and separate calibration with anchor items are compared. Finally, subgroup performance is compared, and the effects of PAA, test-order, student, and school on test scores are described. For results of the test dimensionality within each PAA and across two PAAs, see Yoo, Fu, Wise, and Chung (2011).

Test and Sampling Designs

Table 1 shows the test design of the fall 2009 reading forms. These test forms included two primary PAAs (A and B), with two external linking sets (C1 and C2) embedded into each

PAA to form four PAA forms: PAA-A1, PAA-A2, PAA-B1, and PAA-B2. The external linking items were not used for scoring. Each form included two 50-minute sections. Section I was a scenario-based task set including 20 items focused on either literary skills under a common scenario *Seasons* (Form A) or information/persuasive skills under a common scenario *Wind Power* (Form B). Items were organized under four (Form A) or five (Form B) tasks based on the nature of the questions (e.g., community comments, and solving problems). Section II contained 29 discrete vocabulary items in mini-passage sets including 18 items in Block A or B and 11 external linking items in Block C1 or C2.

Table 1

CBAL Reading Test Design

Section	Items	Description	
		PAA - A1	PAA - A2
I	20	*Seasons*: an extended, integrated scenario-based task set, focused on Literary skills	Same as A1
II	29	Block A (18 items) and external linking Block C1 (11 items): discrete vocabulary items in mini-passage sets focused on Literary and Information/Persuasive skills	Block A (18 items) and external linking Block C2 (11 items)
		PAA - B1	PAA - B2
I	20	*Wind Power*: an extended, integrated scenario-based task set, focused on Information/Persuasive skills	Same as B1
II	29	Block B (18 items) and external linking Block C1 (11 items): discrete vocabulary items in mini-passage sets focused on Literary and Information/Persuasive skills	Block B (18 items) and external linking Block C2 (11 items)

Note. PAA = periodic accountability assessment.

These items measured the content areas of literary, information/persuasive, and vocabulary skills and were classified into three levels in terms of the complexity of skills required by the items in the competency model. Levels 1 and 2 were the two subcategories of model-building skill. Level 1 referred to “identify, retrieve or infer” when activation was high; Level 2 was a more difficult skill and referred to “compare, interpret or infer” when activation

was low. Level 3 was the most difficult skill, applied comprehension (i.e., evaluate, integrate or synthesize).

Each PAA form had both dichotomous and polytomous items. The item types included click-and-click (C&C; i.e., select and copy text from the passage as the answer and paste into the answer box), constructed-response (CR), short CR, and selected-response (SR). Unlike traditional multiple choice items, most of the selected-response items asked examinees to select more than one correct option. Items were either automatically scored by computer or human scored.

Each PAA form had six subscores: Model Building (MB), Applied Comprehension (AC), Information Literacy (IL), Vocabulary (V), Informational (I), and Literary (L). Tables 2–4 list the item information for each item in the two primary PAAs and the two linking sets, including item score ID; the test section, task, and subscore that an item belongs to; item sequence number in the section; item type; scoring type (computer or human scored); and score range. Note that some items are mapped to two subscores. For a description of the test design from a content perspective, see CBAL ELA Team (2011).

The four PAA forms were administered online to a convenience sample of 1,342 Grade 7 students in 17 schools from nine states. See Table 5 for the sample distributions by state, gender, SES (social economic status), ELL (English language learner) status, and race. Each student took two PAAs and was randomly assigned to one of the four test sequences: A1-B2, A2-B1, B1-A2, and B2-A1. Most students (94%) completed both PAAs within one month (mean = 9 days, standard deviation = 12 days).

Classical Item Analyses

Rater Agreement for Human Scored Items

Seasons had two human scored items and *Wind Power* had five human scored items. Each item was rated twice, and the two raters were the same across items and students. A third rater scored an item if the first two raters' scores were not the same. All raters were familiar with the CBAL reading tests. According to the adjudication rule, the final score was the common score of any two raters or the median score when the three raters' scores were not the same. The first two rater scores were used to assess rater agreement. Students receiving an omit or not-reached score on a human scored item were excluded from the analysis of that item.

Table 2

PAA-A: Item and Subscore Information

Task	Item sequence within section	Item score ID	Type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
		Section I (seasons)									
Sound of Summer Running	1	R_SEASONS11	SCR	H	0-1	1					1
	2	R_SEASONS12	SR	A	0-1	1					1
	3	R_SEASONS13	C&C	A	0-2	1					1
	4	R_SEASONS14	CR	H	0-1	1					1
	5	R_SEASONS15	SR	A	0-1	1					1
	6	R_SEASONS16	C&C	A	0-1	1					1
	7	R_SEASONS17	SR	A	0-1	1					1
	8	R_SEASONS18	SR	A	0-1	1					1
	9	R_SEASONS19	SR	A	0-2		1				1
Berkshires in April	10	R_SEASONS21	SR	A	0-2	1					1
	11	R_SEASONS22	SR	A	0-2	1					1
	12	R_SEASONS23	C&C	A	0-2	1					1
	13	R_SEASONS24	C&C	A	0-2	1					1
	14	R_SEASONS25	SCR	A	0-1	1					1
Combined	15	R_SEASONS31	SR	A	0-1		1				1
	16	R_SEASONS32	SR	A	0-1		1				1
Using Rubric	17	R_SEASONS41	SR	A	0-2		1				1
	18	R_SEASONS42	SR	A	0-2		1				1
	19	R_SEASONS43	SR	A	0-2		1				1
	20	R_SEASONS44	SR	A	0-2		1				1
		Block A in Section II									
	1	CBAL_R_Equating_BA02	SR	A	0-1	1					1
	2	CBAL_R_Equating_BA03	SR	A	0-2	1					1
	3	CBAL_R_Equating_BA04	SR	A	0-1				1		
	4	CBAL_R_Equating_BA05	C&C	A	0-1	1				1	
	5	CBAL_R_Equating_BA06	SR	A	0-1	1				1	
	6	CBAL_R_Equating_BA07	SR	A	0-1	1				1	
	7	CBAL_R_Equating_BA08	SR	A	0-1				1		
	8	CBAL_R_Equating_BA09	SR	A	0-1	1				1	
	9	CBAL_R_Equating_BA10	SR	A	0-1	1				1	
	21	CBAL_R_Equating_BA11	SR	A	0-1				1		
	22	CBAL_R_Equating_BA12	SR	A	0-1	1				1	

Task	Item sequence within section	Item score ID	Type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
	23	CBAL_R_Equating_BA13	C&C	A	0-1	1				1	
	24	CBAL_R_Equating_BA14	SR	A	0-1		1			1	
	25	CBAL_R_Equating_BA15	SR	A	0-1				1		
	26	CBAL_R_Equating_BA16	SR	A	0-2	1				1	
	27	CBAL_R_Equating_BA17	SR	A	0-1	1				1	
	28	CBAL_R_Equating_BA18	SR	A	0-1				1		
	29	CBAL_R_Equating_BA01	SR	A	0-1				1		

Note. A = automatically scored by computer, C&C = click & click, = constructed response, H = human scored, PAA = periodic accountability assessment, Subscore S1 = model building (MB), S2 = applied comprehension (AC), S3 = information literacy (IL), S4 = vocabulary (V), S5 = informational (I), S6 = literary (L), SR = selected response, SCR = short CR.

Table 3

PAA-B: Item and Subscore Information

Task	Item sequence within section	Item score ID	Type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
		Section I (Wind Power)									
How Wind Power Works	1	R_WIND_POWER_11	SR	A	0-1	1				1	
	2	R_WIND_POWER_12	SR	A	0-1	1				1	
	3	R_WIND_POWER_13	SR	A	0-2	1				1	
	4	R_WIND_POWER_14	SR	A	0-1		1				
Find Information	5	R_WIND_POWER_21	C&C	A	0-1			1			
	6	R_WIND_POWER_22	SR	A	0-1			1			
	7	R_WIND_POWER_23	SR	A	0-1			1			
	8	R_WIND_POWER_24	SR	A	0-1			1			
Possibilities & Challenges	9	R_WIND_POWER_31	SCR	H	0-1	1				1	
	10	R_WIND_POWER_32	SR	A	0-1			1			
	11	R_WIND_POWER_33	SCR	H	0-2	1				1	
	12	R_WIND_POWER_34	C&C	A	0-2	1				1	
Community Comments	13	R_WIND_POWER_41	C&C	A	0-2			1			
	14	R_WIND_POWER_42	C&C	A	0-1		1				
	15	R_WIND_POWER_43	SR/CR	H	0-2		1			1	

Task	Item sequence within section	Item score ID	Type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
	16	R_WIND_POWER_44	SR/C&C/ CR	H	0-2		1			1	
Solving Problems	17	R_WIND_POWER_51	SR/C&C	A	0-1	1				1	
	18	R_WIND_POWER_52	SR	A	0-1		1			1	
	19	R_WIND_POWER_53	C&C	A	0-1		1			1	
	20	R_WIND_POWER_54	CR	H	0-2	1				1	
			Block B in Section II								
	1	CBAL_R_Equating_BB02	C&C	A	0-1	1					1
	2	CBAL_R_Equating_BB03	SR	A	0-1	1					1
	3	CBAL_R_Equating_BB04	SR	A	0-1				1		
	4	CBAL_R_Equating_BB05	SR	A	0-1	1				1	
	5	CBAL_R_Equating_BB06	SR	A	0-2	1				1	
	6	CBAL_R_Equating_BB07	SR	A	0-2		1			1	
	7	CBAL_R_Equating_BB08	SR	A	0-1				1		
	8	CBAL_R_Equating_BB09	SR	A	0-1	1				1	
	9	CBAL_R_Equating_BB10	SR	A	0-1	1				1	
	21	CBAL_R_Equating_BB11	SR	A	0-1				1		
	22	CBAL_R_Equating_BB12	SR	A	0-1	1				1	
	23	CBAL_R_Equating_BB13	C&C	A	0-1	1				1	
	24	CBAL_R_Equating_BB14	SR	A	0-1	1				1	
	25	CBAL_R_Equating_BB15	SR	A	0-1				1		
	26	CBAL_R_Equating_BB16	SR	A	0-1	1				1	
	27	CBAL_R_Equating_BB17	C&C	A	0-1	1				1	
	28	CBAL_R_Equating_BB18	SR	A	0-1				1		
	29	CBAL_R_Equating_BB01	SR	A	0-1				1		

Note. A = automatically scored by computer, C&C = click & click, = constructed response, H = human scored, PAA = periodic accountability assessment, Subscore S1 = model building (MB), S2 = applied comprehension (AC), S3 = information literacy (IL), S4 = vocabulary (V), S5 = informational (I), S6 = literary (L), SR = selected response, SCR = short CR.

Table 4

Linking Blocks C1 and C2: Item and Subscore Information

Section	Item sequence within section	Item score ID	Type	Scoring type	Score range	S1	S2	S3	S4	S5	S6
Linking Block C1 in Section II	10	CBAL_R_Equating_C05	C&C	A	0-1	1				1	
	11	CBAL_R_Equating_C06	SR	A	0-1	1				1	
	12	CBAL_R_Equating_C09	SR	A	0-1		1	1		1	
	13	CBAL_R_Equating_C10	C&C	A	0-2	1				1	
	14	CBAL_R_Equating_C08	SR	A	0-2		1			1	
	15	CBAL_R_Equating_C01	SR	A	0-1				1		
	16	CBAL_R_Equating_C13	C&C	A	0-2	1					1
	17	CBAL_R_Equating_C15	C&C	A	0-2	1					1
	18	CBAL_R_Equating_C16	C&C	A	0-2	1					1
	19	CBAL_R_Equating_C22	SR	A	0-1		1				1
	20	CBAL_R_Equating_C23	SR	A	0-1		1				1
Linking Block C2 in Section II	10	CBAL_R_Equating_C02	SR	A	0-1	1				1	
	11	CBAL_R_Equating_C03	SR	A	0-1	1				1	
	12	CBAL_R_Equating_C04	C&C	A	0-2	1				1	
	13	CBAL_R_Equating_C07	SR	A	0-1	1		1		1	
	14	CBAL_R_Equating_C11	C&C	A	0-1		1			1	
	15	CBAL_R_Equating_C12	SR	A	0-1				1		
	16	CBAL_R_Equating_C14	C&C	A	0-2	1					1
	17	CBAL_R_Equating_C18	C&C	A	0-1	1					1
	18	CBAL_R_Equating_C19	C&C	A	0-2	1					1
	19	CBAL_R_Equating_C20	C&C	A	0-2		1				1
	20	CBAL_R_Equating_C21	SR	A	0-2		1				1

Note. A = automatically scored by computer, C&C = click & click, = constructed response, H = human scored, PAA = periodic accountability assessment, Subscore S1 = model building (MB), S2 = applied comprehension (AC), S3 = information literacy (IL), S4 = vocabulary (V), S5 = informational (I), S6 = literary (L), SR = selected response, SCR = short CR.

Table 5

Sample Distributions by Demographics

Demographic	*N*	%
State		
Alabama	119	8.9
Arizona	98	7.3
Arkansas	414	30.8
California	68	5.1
Georgia	227	16.9
Kentucky	67	5.0
Louisiana	116	8.6
Massachusetts	101	7.5
Mississippi	84	6.3
Unreported	48	3.6
Gender		
Male	635	47.3
Female	657	49.0
Unreported	50	3.7
Low SES status		
No	361	26.9
Yes	521	38.8
Unreported	460	34.3
ELL status		
No	953	71.0
Yes	45	3.4
Unreported	344	25.6
Race		
African American	332	24.7
Asian/Pacific Islander	39	2.9
Hispanic	45	3.4
Native American	9	.7
White	653	48.7
Unreported	264	19.7

Note. ELL = English language learner, SES = social economic status.

Kappa coefficient and percentage of agreement. Table 6 shows the weighted kappa coefficient for each human scored item as a measure of interrater agreement between the first two raters, the sample size used in each kappa calculation, the asymptotic standard error (ASE) estimate of each weighted kappa coefficient, and the percentage of rater agreement. The weights used for the kappa calculations were the Fleiss-Cohen (1973) weights (commonly known as quadratic weights). The quadratic weight for a pair of raters with score difference d was $1-d^2/k^2$, where k was the score difference between the highest score category and the lowest score category of an item. For dichotomous items, the weighted kappa coefficients were the same as the unweighted kappa coefficients. The weighted kappa coefficient in this case was equivalent to the interclass correlation coefficient as demonstrated in Fleiss and Cohen. The weighted kappa coefficients were in the range of .92 to .96, and the percentage of rater agreement ranged from 80% to 97%. One possible interpretation of kappa is as follows (Altman, 1991, p. 404):

Poor agreement = Less than .20

Fair agreement = .20 to .40

Moderate agreement = .40 to .60

Good agreement = .60 to .80

Very good agreement = .80 to 1.00.

Therefore, all the human scored items showed very good agreement between the first two raters.

Table 6

Weighted Kappa Coefficient and Percentage of Agreement

Human scored item	Effective sample size	Weighted kappa [a]	ASE of kappa	% of agreement
R_SEASONS11	1,218	.89	.01	95.7
R_SEASONS14	1,219	.83	.02	91.4
R_WIND_POWER_31	1,210	.94	.01	97.0
R_WIND_POWER_33	1,206	.96	.01	95.4
R_WIND_POWER_43	1,207	.95	.01	93.0
R_WIND_POWER_44	1,202	.92	.01	91.6
R_WIND_POWER_54	1,173	.82	.01	80.1

Note. ASE = asymptotic standard error.

[a] Quadratic weights (Fleiss & Cohen, 1973).

Generalizability analysis. Generalizability theory (Brennan, 2001; Shavelson & Webb, 1991) was used to estimate rater reliabilities. Treating the human scoring design as a G-study design, we had a balanced design with one facet, rater, and the object of measurement, students. Based on the variance components from the G-studies, we estimated the generalizability coefficients for the following four D-studies (Crocker & Algina, 1986, pp. 157–171):

1. Each student was rated by one rater, and each student had the same rater;
2. Each student was rated by two raters, each student had the same raters, and the final item score was the average of the two rater scores;
3. Each student was rated by one rater, and each student had a different rater;
4. Each student was rated by two raters, each student had different raters, and the final item score was the average of the two rater scores.

Table 7 shows the generalizability coefficient estimates. Two observations could be drawn from Table 7: (a) averaging two rater scores increased the generalizability coefficient estimates of the one-rater model by .02 to .08 across all the human scored items and doubled the accuracy of estimates as indicated by the signal/noise ratios in Table 7; and (b) whether the rater(s) were the same or different across students had little impact on the reliability estimates and their accuracy, because the variances of rater effects in the G-studies were close to 0, which was consistent with the high rater agreement on each item as shown in Table 6. As mentioned above, CBAL reading used different scoring rules to combine two or three rater scores than those in the D-studies with two raters, and the CBAL scoring rules ensured more homogenous rater scores before aggregation. Therefore, the CBAL scoring rules should create scores with higher reliabilities than those from the D-studies with two raters. Given the high single rater reliabilities, the improvement in the reliabilities by using multiple readers may not be substantial.

Item Summary Statistics

Tables 8 and 9 contain item summary statistics for PAA-A and PAA-B, respectively, including: sample size (*N*), mean, standard deviation, maximum possible score point, P+ value, item-total polyserial correlation, item-total Pearson correlation, percentage omit, percentage not reached, percentage system error, and percentage not responding (sum of percentageage omit, not reached, and system error). Tables 10 and 11 contain these item statistics for linking Blocks

C1 and C2, respectively, and four additional statistics: item polyserial and Pearson correlations with the respective total score of PAA-A and PAA-B. At the end of Tables 8–11, mean, standard deviation, minimum, and maximum across all items are included. Tables A1–A3 in Appendix A list the item score frequencies, including the frequencies for omit and not reached responses, as well as system errors (i.e., the online testing system failed to capture a student's response) for PAA-A, PAA-B and linking Blocks C1 and C2, respectively. Note that, unless explicitly specified, item omits were treated as zero, while not reached and system error were treated as missing, and a composite score including any missing item score was assigned to be missing.

Table 7

Generalizability Coefficients for Item Rater D-Studies

Human-scored item	D-study Design 1: 1 rater/same		D-study Design 2: 1 rater/difference		D-study Design 3: 2 rater/same		D-study Design 4: 2 rater/difference	
	G coeff.	Signal /noise ratio	G coeff.	Signal /noise ratio	G coeff.	Signal /noise ratio	G coeff.	Signal /noise ratio
R_SEASONS11	.89	8.06	.94	16.11	.89	8.04	.94	16.08
R_SEASONS14	.84	5.09	.91	10.19	.83	4.85	.91	9.70
R_WIND_POWER_31	.94	15.88	.97	31.76	.94	15.87	.97	31.74
R_WIND_POWER_33	.96	26.15	.98	52.29	.96	26.14	.98	52.29
R_WIND_POWER_43	.95	18.32	.97	36.63	.95	17.49	.97	34.98
R_WIND_POWER_44	.92	12.20	.96	24.41	.92	12.07	.96	24.14
R_WIND_POWER_54	.83	4.93	.91	9.87	.82	4.60	.90	9.20

Tables 8–11 show that the percentages of missing item responses were small, as the maximum percentage of students not responding for an item was 3.21%. Also presented in the table are P+ values, which refer to the proportion of correct responses for an item. For a dichotomous item, the P+ value is the same as the mean, while for a polytomous item, the P+ statistic is calculated as the ratio of the mean to the maximum possible score. The item P+ values for PAA-A, PAA-B, and linking Blocks C1 and C2 were between .08 and .78 with the averages .45, .46, .40, and .46, respectively, which indicates that PAA-A, PAA-B, and the linking sets had similar difficulties.

Tables 8–11 also give item-total correlations. The correlation between an item score and the total score is used to indicate the strength of the association between an item and the construct that it presumably measures (represented by total score), which is closely related to test

Table 8

PAA-A: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,145)	Pearson correlation (*N* = 1,145)	% omit	% not reached	% system error	% no response
R_SEASONS11	1,223	.73	.45	1	.73	.73	.52	.41	0	0	.41
R_SEASONS12	1,223	.30	.46	1	.30	.63	.49	0	0	0	0
R_SEASONS13	1,223	.90	.82	2	.45	.63	.56	0	0	0	0
R_SEASONS14	1,223	.50	.50	1	.50	.53	.42	.33	0	0	.33
R_SEASONS15	1,222	.26	.44	1	.26	.52	.40	0	.08	0	.08
R_SEASONS16	1,221	.34	.47	1	.34	.44	.35	0	.16	0	.16
R_SEASONS17	1,221	.42	.49	1	.42	.58	.47	0	.16	0	.16
R_SEASONS18	1,221	.28	.45	1	.28	.54	.41	0	.16	0	.16
R_SEASONS19	1,221	.94	.89	2	.47	.64	.57	0	.16	0	.16
R_SEASONS21	1,220	1.26	.89	2	.63	.65	.55	0	.25	0	.25
R_SEASONS22	1,220	.68	.67	2	.34	.30	.27	0	.25	0	.25
R_SEASONS23	1,220	.88	.70	2	.44	.50	.45	0	.25	0	.25
R_SEASONS24	1,220	1.32	.65	2	.66	.40	.35	0	.25	0	.25
R_SEASONS25	1,219	.31	.46	1	.31	.70	.54	0	.33	0	.33
R_SEASONS31	1,217	.42	.49	1	.42	.45	.36	0	.49	0	.49
R_SEASONS32	1,216	.42	.49	1	.42	.73	.59	0	.57	0	.57
R_SEASONS41	1,211	1.03	.93	2	.52	.35	.30	0	.90	.08	.98
R_SEASONS42	1,206	.69	.78	2	.35	.26	.23	0	1.39	0	1.39
R_SEASONS43	1,203	.90	.91	2	.45	.57	.51	0	1.64	0	1.64
R_SEASONS44	1,203	.59	.82	2	.29	.44	.38	0	1.64	0	1.64
CBAL_R_Equating_BA01	1,175	.65	.48	1	.65	.41	.31	0	.17	0	.17
CBAL_R_Equating_BA02	1,175	.62	.49	1	.62	.64	.49	0	0	.17	.17
CBAL_R_Equating_BA03	1,176	1.14	.81	2	.57	.69	.62	0	0	.08	.08
CBAL_R_Equating_BA04	1,177	.61	.49	1	.61	.62	.48	0	0	0	0
CBAL_R_Equating_BA05	1,177	.26	.44	1	.26	.45	.34	0	0	0	0
CBAL_R_Equating_BA06	1,177	.38	.49	1	.38	.32	.26	0	0	0	0
CBAL_R_Equating_BA07	1,177	.69	.46	1	.69	–	.57	0	0	0	0
CBAL_R_Equating_BA08	1,177	.60	.49	1	.60	.63	.49	0	0	0	0
CBAL_R_Equating_BA09	1,177	.36	.48	1	.36	.29	.23	0	0	0	0
CBAL_R_Equating_BA10	1,176	.45	.50	1	.45	.66	.53	0	.08	0	.08

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,145)	Pearson correlation (*N* = 1,145)	% omit	% not reached	% system error	% no response
CBAL_R_Equating_BA11	1,175	.27	.45	1	.27	.53	.41	0	.08	.08	.17
CBAL_R_Equating_BA12	1,176	.08	.28	1	.08	.28	.17	0	.08	0	.08
CBAL_R_Equating_BA13	1,175	.54	.50	1	.54	.67	.53	0	.08	.08	.17
CBAL_R_Equating_BA14	1,176	.34	.47	1	.34	.36	.29	0	.08	0	.08
CBAL_R_Equating_BA15	1,176	.52	.50	1	.52	.67	.53	0	.08	0	.08
CBAL_R_Equating_BA16	1,176	.85	.79	2	.43	.64	.59	0	.08	0	.08
CBAL_R_Equating_BA17	1,176	.39	.49	1	.39	.68	.54	0	.08	0	.08
CBAL_R_Equating_BA18	1,176	.73	.44	1	.73	.72	.50	0	.08	0	.08
Mean		.60	.57	1.32	.45	.54	.44	.02	.25	.01	.29
SD		.30	.17	.47	.15	.14	.12	.08	.43	.04	.42
Min		.08	.28	1	.08	.26	.17	0	0	0	0
Max		1.32	.93	2	.73	.73	.62	.41	1.64	.17	1.64

Note. PAA = periodic accountability assessment.

[a] Missing cell = item-total polyserial correlation did not converge.

Table 9

PAA - B: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,116)	Pearson correlation (*N* = 1,116)	% omit	% not reached	% system error	% no response
R_WIND_POWER_11	1,214	.42	.49	1	.42	.61	.49	0	0	0	0
R_WIND_POWER_12	1,214	.46	.50	1	.46	.75	.60	0	0	0	0
R_WIND_POWER_13	1,214	.90	.92	2	.45	.60	.51	0	0	0	0
R_WIND_POWER_14	1,214	.38	.49	1	.38	.32	.26	0	0	0	0
R_WIND_POWER_21	1,214	.40	.49	1	.40	.30	.24	0	0	0	0
R_WIND_POWER_22	1,214	.53	.50	1	.53	.63	.50	0	0	0	0
R_WIND_POWER_23	1,214	.35	.48	1	.35	.18	.14	0	0	0	0
R_WIND_POWER_24	1,214	.55	.50	1	.55	.47	.37	0	0	0	0
R_WIND_POWER_31	1,213	.47	.50	1	.47	–	.56	.25	.08	0	.33
R_WIND_POWER_32	1,213	.57	.50	1	.57	.64	.50	0	.08	0	.08
R_WIND_POWER_33	1,212	.92	.81	2	.46	.72	.66	.41	.16	0	.58
R_WIND_POWER_34	1,211	1.08	.75	2	.54	.70	.64	0	.25	0	.25
R_WIND_POWER_41	1,210	.92	.73	2	.46	–	.56	0	.33	0	.33
R_WIND_POWER_42	1,207	.18	.38	1	.18	.38	.27	0	.49	.08	.58
R_WIND_POWER_43	1,208	.83	.82	2	.42	.58	.53	.08	.49	0	.58
R_WIND_POWER_44	1,205	.46	.77	2	.23	.78	.64	.25	.74	0	.99
R_WIND_POWER_51	1,198	.19	.39	1	.19	.61	.44	0	1.32	0	1.32
R_WIND_POWER_52	1,192	.09	.28	1	.09	.72	.42	0	1.81	0	1.81
R_WIND_POWER_53	1,189	.41	.49	1	.41	.65	.52	0	2.06	0	2.06
R_WIND_POWER_54	1,184	.72	.84	2	.36	.80	.73	.74	2.47	0	3.21
CBAL_R_Equating_BB01	1,152	.51	.50	1	.51	.36	.28	0	.09	0	.09
CBAL_R_Equating_BB02	1,153	.53	.50	1	.53	.78	.62	0	0	0	0
CBAL_R_Equating_BB03	1,153	.45	.50	1	.45	.57	.46	0	0	0	0
CBAL_R_Equating_BB04	1,153	.78	.41	1	.78	.74	.49	0	0	0	0
CBAL_R_Equating_BB05	1,153	.57	.50	1	.57	.62	.48	0	0	0	0
CBAL_R_Equating_BB06	1,153	.78	.88	2	.39	.70	.62	0	0	0	0
CBAL_R_Equating_BB07	1,153	1.16	.81	2	.58	–	.54	0	0	0	0
CBAL_R_Equating_BB08	1,153	.56	.50	1	.56	.42	.34	0	0	0	0
CBAL_R_Equating_BB09	1,153	.59	.49	1	.59	.60	.47	0	0	0	0
CBAL_R_Equating_BB10	1,153	.50	.50	1	.50	.60	.48	0	0	0	0

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,116)	Pearson correlation (*N* = 1,116)	% omit	% not reached	% system error	% no response
CBAL_R_Equating_BB11	1,152	.64	.48	1	.64	.66	.50	0	.09	0	.09
CBAL_R_Equating_BB12	1,152	.42	.49	1	.42	–	.56	0	.09	0	.09
CBAL_R_Equating_BB13	1,152	.65	.48	1	.65	.64	.48	0	.09	0	.09
CBAL_R_Equating_BB14	1,151	.30	.46	1	.30	.25	.19	0	.09	.09	.17
CBAL_R_Equating_BB15	1,152	.59	.49	1	.59	.59	.46	0	.09	0	.09
CBAL_R_Equating_BB16	1,152	.34	.48	1	.34	.59	.47	0	.09	0	.09
CBAL_R_Equating_BB17	1,152	.40	.49	1	.40	.85	.68	0	.09	0	.09
CBAL_R_Equating_BB18	1,152	.66	.47	1	.66	.63	.47	0	.09	0	.09
Mean		.56	.55	1.24	.46	.59	.48	.05	.29	0	.34
SD		.24	.16	.43	.14	.16	.14	.14	.6	.02	.69
Min		.09	.28	1	.09	.18	.14	0	0	0	0
Max		1.16	.92	2	.78	.85	.73	.74	2.47	.09	3.21

Note. PAA = periodic accountability.

[a] Missing cells = item-total polyserial correlation did not converge.

Table 10

Linking Block C1: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,170)	Polyserial A[b] (*N* = 1,085)	Polyserial B[c] (*N* = 1,056)	Pearson corr.[a] (*N* = 1,170)	Pearson corr. A[c] (*N* = 1,085)	Pearson corr. B[c] (*N* = 1,056)	% omit	% not reached	% system error	% no response
CBAL_R_Equating_C05	1,171	.59	.49	1	.59	.71	.59	.60	.56	.46	.47	0	.09	0	.09
CBAL_R_Equating_C06	1,171	.40	.49	1	.40	.38	.23	.28	.31	.18	.22	0	.09	0	.09
CBAL_R_Equating_C09	1,171	.16	.37	1	.16	.11	-.02	.02	.07	-.01	.01	0	.09	0	.09
CBAL_R_Equating_C10	1,171	1.43	.76	2	.72	.81	.65	–[d]	.66	.52	.53	0	.09	0	.09
CBAL_R_Equating_C08	1,171	.85	.85	2	.42	.70	.52	.58	.63	.47	.52	0	.09	0	.09
CBAL_R_Equating_C01	1,171	.47	.50	1	.47	.58	.47	.49	.46	.37	.40	0	.09	0	.09
CBAL_R_Equating_C13	1,171	.82	.66	2	.41	.53	.39	.40	.48	.35	.36	0	.09	0	.09
CBAL_R_Equating_C15	1,171	.46	.81	2	.23	.78	.61	.60	.61	.48	.46	0	.09	0	.09
CBAL_R_Equating_C16	1,171	1.01	.91	2	.51	.84	.71	.71	.76	.63	.63	0	.09	0	.09
CBAL_R_Equating_C22	1,170	.27	.44	1	.27	.61	.51	.46	.47	.39	.36	0	.09	.09	.17
CBAL_R_Equating_C23	1,171	.27	.44	1	.27	.75	.63	.61	.57	.47	.46	0	.09	0	.09
Mean		.61	.61	1.45	.40	.62	.48	.48	.51	.39	.40	0	.09	.01	.09
SD		.38	.19	.52	.17	.22	.21	.20	.19	.17	.17	0	0	.03	.03
Min		.16	.37	1	.16	.11	-.02	.02	.07	-.01	.01	0	.09	0	.09
Max		1.43	.91	2	.72	.84	.71	.71	.76	.62	.63	0	.09	.09	.17

Note. corr = correlation.

[a] Polyserial or Pearson item correlation with the total score of Block C1. [b] Polyserial or Pearson item correlation with the total score of PAA-A. [c] Polyserial or Pearson item correlation with the total score of PAA-B. [d] Polyserial item-total correlation did not converge.

Table 11

Linking Block C2: Item Statistics

Item score ID	*N*	Mean	SD	Max possible score	P+	Polyserial[a] (*N* = 1,155)	Polyserial A[b] (*N* = 1,070)	Polyserial B[c] (*N* = 1,061)	Pearson corr.[a] (*N* = 1,155)	Pearson corr. A[b] (*N* = 1,070)	Pearson corr. B[c] (*N* = 1,061)	% omit	% not reached	% system error	% no response
CBAL_R_Equating_C02	1,158	.77	.42	1	.77	.85	.66	.72	.55	.45	.49	0	0	0	0
CBAL_R_Equating_C03	1,158	.50	.50	1	.50	.77	.64	.62	.61	.51	.50	0	0	0	0
CBAL_R_Equating_C04	1,157	.69	.91	2	.35	.80	.62	.65	.68	.52	.55	0	0	.09	.09
CBAL_R_Equating_C07	1,158	.67	.47	1	.67	.62	.49	.47	.47	.37	.36	0	0	0	0
CBAL_R_Equating_C11	1,158	.28	.45	1	.28	.50	.39	.42	.39	.31	.33	0	0	0	0
CBAL_R_Equating_C12	1,158	.29	.46	1	.29	.54	.41	.36	.42	.32	.28	0	0	0	0
CBAL_R_Equating_C14	1,157	.95	.74	2	.47	.48	.29	.28	.43	.27	.25	0	0	.09	.09
CBAL_R_Equating_C18	1,158	.47	.50	1	.47	.77	.67	.64	.61	.54	.52	0	0	0	0
CBAL_R_Equating_C19	1,158	.94	.92	2	.47	.84	.70	.70	.76	.61	.62	0	0	0	0
CBAL_R_Equating_C20	1,157	.93	.80	2	.46	.70	.57	.57	.64	.52	.52	0	0	.09	.09
CBAL_R_Equating_C21	1,158	.56	.81	2	.28	.74	–[d]	.55	.64	.52	.48	0	0	0	0
Mean		.64	.63	1.45	.46	.69	.54	.55	.56	.45	.44	0	0	.02	.02
SD		.24	.20	.52	.16	.13	.14	.14	.12	.11	.12	0	0	.04	.04
Min		.28	.42	1	.28	.48	.29	.28	.39	.26	.25	0	0	0	0
Max		.95	.92	2	.77	.85	.70	.72	.76	.61	.62	0	0	.09	.09

Note. corr = correlation.

[a] Polyserial or Pearson item correlation with the total score of Block C2. [b] Polyserial or Pearson item correlation with the total score of PAA-A. [c] Polyserial or Pearson item correlation with the total score of PAA-B. [d] Polyserial item-total correlation did not converge.

reliability. In this case, the polyserial correlation is preferred to the ordinary Pearson correlation, because the polyserial correlation more closely reflects the actual relationship between an ordinal variable and a continuous variable, while the Pearson correlation tends to underestimate this relationship. The polyserial correlation assumes that the ordinal variable has an underlying standard normal distribution, and the two variables follow a bivariate normal distribution. Tables 8–11 include both polyserial and Pearson item-total correlations because some polyserials did not converge. All polyserials were higher than their Pearson correlation counterparts. The mean item-total polyserial correlations for PAA-A and PAA-B were .54 and .59, respectively, and the mean item polyserial correlations with the respective total scores of PAA-A and PAA-B in the linking Blocks C1 and C2 were between .48 and .55. All item-total correlations look reasonable except for item CBAL_R_Equating_C09 in Block C1, which had correlations -.02 and .02 with the total scores of PAA-A and PAA-B, respectively. Consequently, this item was removed from subsequent analyses.

Table 12 shows the average item P+ values by required skill level, where that designation refers to the categorization in terms of the CBAL reading competency model. Over all items collapsing across PAAs (including linking items), as well as for the items on PAA-B alone, the average item P+ values decreased as item skill level increased, a result theoretically in keeping with the CBAL competency model categorizations. However, for PAA-A, the average item P+ values increased from skill Level 2 to Level 3, an inconsistent result, which could suggest that the classification of items in terms of skill level could be refined.

Table 12

Average Item P+ Value by Item Skill Level

Test form	Level 1		Level 2		Level 3	
	Number of items	Mean P+ (max. *N*)	Number of items	Mean P+ (max. *N*)	Number of items	Mean P+ (max. *N*)
PAA-A	9	.52 (1,223)	14	.36 (1,223)	9	.43 (1,221)
PAA-B	8	.49 (1,214)	10	.46 (1,214)	14	.37 (1,214)
All (PAA-A + PAA-B + linking sets C1 and C2)	22	.52 (1,223)	32	.42 (1,223)	30	.38 (1,221)

Note. PAA = periodic accountability assessment.

Differential Item Functioning

Test fairness requires that all test items function in the same way for students from different population groups. Differential item functioning (DIF) analysis is used to identify items that may have such biases. An item shown to have DIF may be measuring some construct different from what it intends to measure. For such an item, further review and/or revision by content experts is needed. In this study, the Mantel-Haenszel procedure was used to detect DIF (Dorans & Holland, 1993; Holland & Thayer, 1988; Zwick, Donoghue, & Grima, 1993). ETS's DIF classification includes three DIF categories: A, B, and C (Dorans & Holland, 1993). Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of DIF. In practice, only Category C items are considered to require further review/revision. The DIF analyses were conducted for the following demographic group pairs:

1. Gender (Male vs. Female)
2. Race/Ethnicity (White vs. Black; White vs. combination of Native American, Asian/Pacific Islander, and Hispanic)
3. Low SES students (No vs. Yes)

Table 13 lists the Category C DIF items; the tables in Appendix B show the DIF category for every item. Three items were found to have Category C DIF: one item in *Wind Power* and two items in the linking blocks. Note that the ethnicity combination group had sample sizes smaller than 200; therefore, their DIF results should be interpreted with caution.

Table 13

Category C DIF Items

Item score ID	C DIF description
R_WIND_POWER_12	Favor male over female
CBAL_R_Equating_C03	Favor ethnicity combination over White in PAA-B
CBAL_R_Equating_C08	Favor male over female in PAA-B
	Favor White over ethnicity combination in both PAA-A and PAA-B

Note. DIF = differential item functioning. PAA = periodic accountability assessment.

Statistics for Test Section Scores, Subscores and Total Scores

In this section we present the summary statistics, reliabilities, and correlations for test section scores, subscores, and total raw scores. Table 14 shows these statistics for the section raw scores of the linking block (C1 or C2), the operational discrete item block (A or B), Section I (*Seasons* or *Wind Power*), and the main PAA form (PAA-A or PAA-B) within each of the four PAA forms (PAA-A1, PAA-A2, PAA-B1, and PAA-B2). The linking blocks had relatively strong relationships with the operational forms: the correlations of those blocks with the operational discrete item blocks, Section I, and the main PAA forms ranged from .73 to .85. The operational discrete item blocks also had high correlations with the Section I scenario-based task set, running from .74 to .80.

Tables 15 and 16 present the statistics for the subscores and total scores of PAA-A and PAA-B, respectively. One can see that for both PAAs the intersubscore correlations for subscores with mutually exclusive items were between .42 and .75, and the correlations of subscores with total scores were between .65 and .98. The correlation between PAA-A and PAA-B total scores was .79. The reliabilities (standardized Cronbach's alpha) for subscores were between .48 and .89, and the reliabilities for PAA-A and PAA-B were .89 and .91, respectively. The reliabilities based on task scores for PAA-A and PAA-B (i.e., all discrete items were treated as one task set in each PAA), commonly known as testlet reliabilities, were .86 and .89, respectively. The fact that they were very close to the reliabilities based on item scores suggests that the testlet effects at the task level were minor for both PAAs.

Tables 17 and 18 show the statistics for the task scores and total scores of *Seasons* and *Wind Power*, respectively. The intertask correlations were low to moderate in the range between .33 and .64, and the task-total correlations were moderate to high from .60 to .87. The correlation between the scenario-based task set total scores for *Seasons* and *Wind Power* was .69. The reliabilities for task scores (within these task sets) were between .31 and .73, and the reliabilities for the *Seasons* and *Wind Power* task sets were .80 and .84, respectively. The testlet reliabilities for *Seasons* and *Wind Power* were .74 and .83, respectively. The similarity of the testlet reliabilities and the regular reliabilities based on item scores suggests that task did not appear to have strong testlet effects in either *Seasons* or *Wind Power*.

In Tables 14–18, most score means were a little smaller than half of their possible maximum scores, indicating the tests were difficult for the samples assessed.

Table 14

Test Section Score Summary and Correlations Within Each PAA

Score	Number of items	Max poss. score	*N*	Mean	SD	Pearson correlation			
						Block C1	Block A	Seasons	PAA-A
PAA-A1									
Block C1	10	15	607	6.59	3.70	1.00	–	–	–
Block A	18	20	605	9.48	4.43	.79	1.00	–	–
Seasons	20	30	619	13.12	5.82	.73	.78	1.00	–
PAA-A (Seasons + Block A)	38	50	593	22.79	9.66	.80	.93	.96	1.00
PAA-A2									
Block C2	11	16	567	7.03	4.01	1.00	–	–	–
Block A	18	20	566	9.53	4.47	.81	1.00	–	–
Seasons	20	30	583	13.22	5.89	.74	.74	1.00	–
PAA-A (Seasons + Block A)	38	50	552	22.82	9.67	.83	.91	.95	1.00
PAA-B1									
Block C1	10	15	563	6.56	3.63	1.00	–	–	–
Block B	18	20	564	10.33	4.94	.83	1.00	–	–
Wind Power	20	27	572	10.73	6.19	.77	.78	1.00	–
PAA-B (Wind Power + Block B)	38	47	543	21.21	10.48	.85	.93	.95	1.00
PAA-B2									
Block C2	11	16	588	7.05	4.19	1.00	–	–	–
Block B	18	20	587	10.56	4.86	.82	1.00	–	–
Wind Power	20	27	612	10.96	6.15	.80	.80	1.00	–
PAA-B (Wind Power + Block B)	38	47	573	21.58	10.49	.85	.94	.96	1.00

Note. PAA = periodic accountability assessment.

Table 15

PAA-A: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Standardized alpha	Pearson correlation[b]					
							S1	S2	S4	S5	S6	Total
Subscore 1	24	31	1,163	14.01	6.27	.84	1.00	–	–	–	–	–
Subscore 2	8	13	1,150	5.35	2.93	.58	.63	1.00	–	–	–	–
Subscore 4	6	6	1,174	3.39	1.66	.61	.69	.51	1.00	–	–	–
Subscore 5	10	10	1,175	4.35	2.48	.64	.83	.61	.62	1.00	–	–
Subscore 6	22	33	1,148	15.01	6.50	.83	.93	.82	.66	.70	1.00	–
Total	38	50	1,145	22.80	9.66	.89	.96	.80	.78	.84	.97	1.00

Note. PAA = periodic accountability assessment.

[a] Subscore S1 = Model Building (MB), S2 = Applied Comprehension (AC), S4 = Vocabulary (V), S5 = Informational (I), and S6 = Literary (L). [b] Italicized correlations contain items that are mapped to both subscores.

Table 16

PAA-B: Test Subscore and Total Score Summary and Correlations

Score[a]	Number of items	Max poss. score	*N*	Mean	SD	Standardized Alpha	*Pearson correlation[b]*						
							S1	S2	S3	S4	S5	S6	Total
Subscore 1	19	24	1,116	10.78	6.22	.87	1.00	–	–	–	–	–	–
Subscore 2	6	10	1,123	3.54	2.35	.62	.75	1.00	–	–	–	–	–
Subscore 3	6	7	1,210	3.31	1.70	.48	.65	.56	1.00	–	–	–	–
Subscore 4	6	6	1,152	3.74	1.61	.57	.67	.52	.49	1.00	–	–	–
Subscore 5	21	22	1,116	12.78	7.40	.89	*.98*	*.85*	.66	.66	1.00	–	–
Subscore 6	2	2	1,153	.98	.83	.56	*.66*	.50	.42	.46	.58	1.00	–
Total	38	47	1,116	21.40	10.48	.91	.97	.84	.75	.74	.98	.65	1.00

Note. PAA = periodic accountability assessment.

[a] Subscore S1 = Model Building (MB), S2 = Applied Comprehension (AC), S3 = Information Literacy (IL), S4 = Vocabulary (V), S5 = Informational (I), and S6 = Literary (L). [b] Italicized correlations contain items that are mapped to both subscores.

Table 17

Seasons: Task Score and Total Score Summary and Correlations

Score	Number of items	Max poss. score	*N*	Mean	SD	Standardized alpha	*Pearson correlation*				
							Task 1	Task 2	Task 3	Task 4	Total
Task 1	9	11	1,221	4.67	2.86	.73	1.00	–	–	–	–
Task 2	5	9	1,219	4.45	1.98	.51	.55	1.00	–	–	–
Task 3	2	2	1,216	.84	.76	.32	.44	.40	1.00	–	–
Task 4	4	8	1,202	3.22	1.97	.31	.39	.33	.37	1.00	–
Total	20	30	1,202	13.17	5.85	.80	.87	.78	.61	.69	1.00

Table 18

Wind Power: Task Score and Total Score Summary and Correlations

Score	Number of items	Max poss. score	*N*	Mean	SD	Standardized alpha	Pearson correlation					
							Task 1	Task 2	Task 3	Task 4	Task 5	Total
Task 1	4	5	1214	2.16	1.59	.52	1.00	–	–	–	–	–
Task 2	4	4	1214	1.83	1.12	.31	.34	1.00	–	–	–	–
Task 3	4	6	1211	3.04	1.87	.69	.55	.40	1.00	–	–	–
Task 4	4	7	1205	2.39	1.83	.54	.50	.38	.61	1.00	–	–
Task 5	4	5	1184	1.41	1.43	.61	.53	.39	.64	.63	1.00	–
Total	20	27	1184	10.85	6.16	.84	.76	.60	.85	.83	.82	1.00

Table 19 shows the correlations of PAA-A and PAA-B with scores from state tests in English language arts (ELA), math, reading, and writing. Within each subject, PAA-A and PAA-B had very similar correlations on average with the state tests, as shown in the last two rows of Table 19. Across both PAAs, CBAL reading tests had mean correlations of .67 and .70 with state ELA and reading tests, respectively, which were higher than the mean correlation of .60 with state math tests. The relatively strong correlations between both PAAs and ELA and reading state tests, and the fact that their association was stronger than with state math tests, provide some supportive evidence for the validity of CBAL reading tests. Only one school in Arizona provided writing state test scores, and the mean correlation was .53 across both PAAs. Please note the limited sample sizes used in calculating these correlations: except for Arkansas and Georgia, the other four states all had sample sizes of fewer than 100.

Item Response Theory Item Calibration and Scaling

Yoo et al. (2011) showed that the CBAL reading tests employed in the current study had a unidimensional structure both within PAA and across PAAs. Therefore, the CBAL reading tests were calibrated using the unidimensional generalized partial credit model (GPCM; Muraki, 1992). Two calibration approaches, concurrent calibration and separate calibration, were carried out, and the resulting item parameter estimates and ability (theta) estimates were compared. Note that most students took the two tests within a short time period; thus, the two groups taking the two tests were actually the same. Theoretically, we do not need to do equating after separate calibrations. However, because there were linking items, and we would like to check the quality of the linking set, we conducted the separate calibration with anchor linking and compared the estimates to those from the concurrent estimation.

Table 19

Correlations of PAA-A and PAA-B With State Tests

State	CBAL score	*Pearson correlation*			
		ELA (*N*)	Math (*N*)	Reading (*N*)	Writing (*N*)
Alabama (School A) [a]	PAA-A	.62 (47)	.53 (48)	.74 (48)	–
	PAA-B	.62 (47)	.47 (48)	.75 (48)	–
Alabama (School B) [a]	PAA-A	.81 (62)	.74 (62)	.81 (62)	–
	PAA-B	.84 (61)	.78 (61)	.84 (61)	–
Arizona	PAA-A	–	.64 (96)	.71 (96)	.51 (96)
	PAA-B	–	.72 (94)	.79 (94)	.54 (94)
Arkansas	PAA-A	.60 (326)	.59 (328)	–	–
	PAA-B	.61 (322)	.61 (324)	–	–
California	PAA-A	.73 (53)	.46 (52)	.55 (53)	–
	PAA-B	.55 (56)	.29 (55)	.43 (56)	–
Georgia	PAA-A	.58 (204)	.58 (204)	.66 (147)	–
	PAA-B	.62 (202)	.66 (202)	.71 (144)	–
Mississippi	PAA-A	.70 (34)	.61 (34)	–	–
	PAA-B	.79 (28)	.69 (28)	–	–
Mean [b]	PAA-A	.67	.59	.69	.51
	PAA-B	.67	.60	.70	.54

Note. PAA = periodic accountability assessment. ELA = English language arts.

[a] Schools A and B reported test scores on different scales; therefore, their correlations were calculated separately. [b] The simple average of correlations.

In this study, the GPCM was formulized as the following:

$$P_{ijm} = p(x_{ij} = m \mid \theta_j, a_i, \mathbf{b}_i) = \frac{\exp(b_{im} + a_i\theta_j m)}{\sum_{v=0}^{M_i - 1} \exp(b_{iv} + a_i\theta_j v)}$$

where

$b_{i0} \equiv 0$;

x_{ij} is examinee j's score on item i;

M_i is item i's number of score categories;

m is item i's possible integer score point equal to 0 to $M_i - 1$;

b_{im} is the intercept parameter on item *i* for score m;

$\mathbf{b}_i$ is the vector with elements b_{im};

a_i is the discrimination (slope) parameter for item i;

θ_j is examinee j's latent (theta) score, and

P_{ijm} is examinee j's probability of achieving score m on item i.

Note that in the traditional GPCM formulization (Muraki, 1992), b_{im} is written in a summation form as $b_{im} = -a_i \sum_{g=1}^{m} b_{ig}$.

Concurrent Versus Separate Calibrations

In the concurrent calibration, all the items were calibrated together, and for the test forms that a student did not take, item responses were treated as missing in estimating item parameters. However, in the separate calibration, each primary PAA (PAA-A or PAA-B) plus all the linking items (Blocks C1 and C2) were calibrated separately, and then PAA-B item parameter estimates were equated to the PAA-A scale via the linking items. Because the linking items were external ones, theta estimations were based on PAA-A or PAA-B items only. In the separate calibration,

theta estimates in PAA-B were also adjusted to the scale of PAA-A. The expected a posteriori (EAP) method was used to estimate theta. Table 20 shows the sample sizes used in the item calibration and EAP theta estimation for each calibration. Note that a student with any missing value in PAA-A or PAA-B was excluded from the theta estimation for that test.

Table 20

Sample Sizes Used in IRT Calibrations

	PAA-A	PAA-B	PAA-A + Blocks C1 and C2	PAA-B + Blocks C1 and C2	Concurrent calibration
Item parameter estimation	NA	NA	1,229	1,222	1,342
EAP theta estimation	1,145	1,116	NA	NA	NA

Note. EAP = expected a posterior,. IRT = item response theory.

In the separate calibration, the Stocking-Lord test characteristic curve (TCC) method (Stocking & Lord, 1983) was used to link PAA-B to PAA-A through the external linking Blocks C1 and C2. The linking functions for PAA-B were:

$$a_i^* = a_i / A, \text{ and}$$

$$b_{is_{im}}^* = b_{is_{im}} - a_i mB / A,$$

where the superscript asterisk indicated the transformed item parameters, and A and B were the linking constants equal to 1.05 and -.04, respectively. The weighted root mean squared errors of item characteristic curves for the anchor items based on the item parameters from PAA-A and the transformed item parameters from PAA-B were all smaller than .12, indicating no item drift occurred. The weights were from the standard normal distribution with equal intervals of .1 between -4 and 4. Figure 1 compares item discrimination parameter estimates (a_i) and item intercept parameter estimates (b_{i1} and b_{i2}) between the concurrent calibration and separate calibration after anchor linking for PAA-A and PAA-B. For the two PAAs, the Pearson correlations of slopes and intercepts were .99 and 1.00.

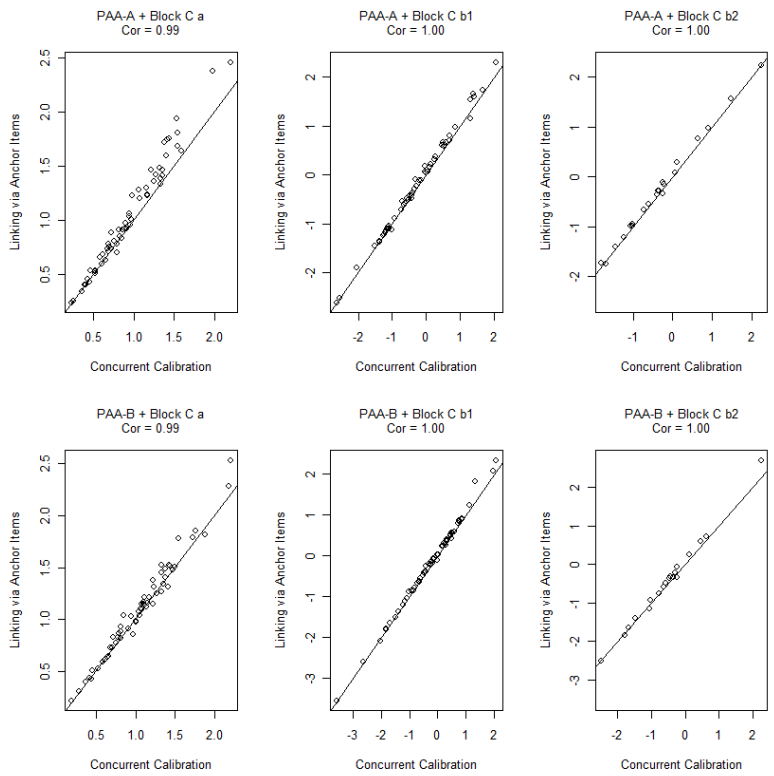

***Figure 1*. Comparison of item parameters between concurrent calibrations and anchor linking. The lines in the plots are diagonal lines. Cor = correlation, PAA = periodic accountability assessment.**

Figure 2 shows the comparisons of EAP theta estimates between the two calibration approaches for PAA-A and PAA-B. The EAP theta estimates in PAA-B in the separate calibration were rescaled to the scale of PAA-A by:

$$\theta_j^* = A\theta_j + B,$$

where A and B were the same linking constants used above for item parameter transformation. The EAP thetas from the separate calibration were adjusted to have the same population mean and standard deviation of the combined two PAAs as the ones from the concurrent calibration, so that they were on the same metric. The Pearson correlations of EAP thetas for both PAAs were 1 after rounding, and the theta estimates were very similar between the two calibrations, as indicated by the small root mean square errors (RMSEs) of .04 and .03.

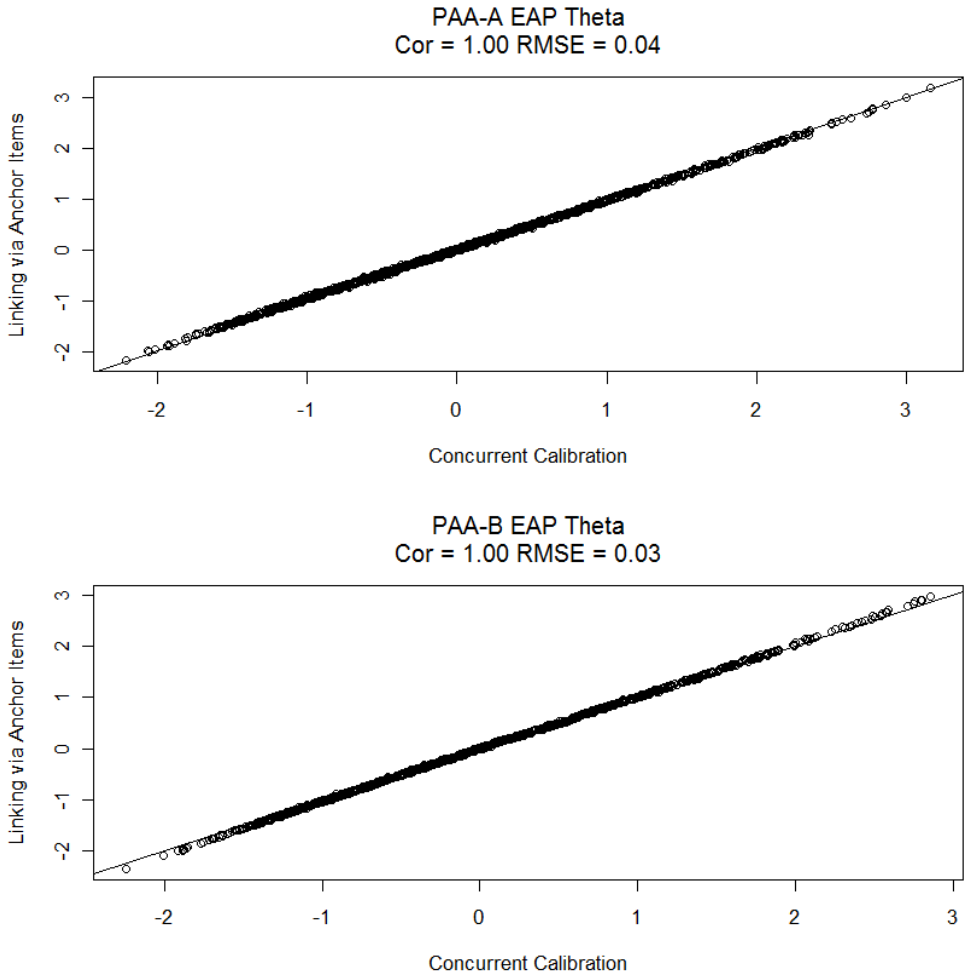

***Figure 2*. Comparison of EAP thetas between concurrent calibrations and anchor linking. The lines in the plots are diagonal lines. Cor = correlation, EAP = expected a posteriori, PAA = periodic accountability assessment, RMSE = root mean square error.**

In conclusion, item parameter and EAP theta estimates from both calibrations were almost perfectly correlated, and RMSEs for the EAP theta estimates were quite small. Therefore, the results from the two calibrations were interchangeable, especially because in practice theta estimates are often of the ultimate concern. In the following section, the results from the concurrent calibration are reported.

Item Parameter and Theta Estimates

Tables 21–23 list the item parameter estimates, standard errors, and significance of item chi-square fit statistics for PAA-A, PAA-B, and the linking blocks. All item parameter estimates were within a reasonable range, and the means of item discrimination parameter estimates (a_i) and item intercept parameter estimates (b_{i1} and b_{i2}) were .90, -.24, and -.24 for PAA-A; 1.07, -.22, and -.61 for PAA-B; and 1.05, -.34, and -.47 for the linking sets. The means of item parameter estimates of PAA-A, PAA-B, and linking blocks were similar, indicating that these three test forms/item blocks performed similarly, which is particularly important for the linking sets, as this is a requirement for such a set. Because the means of item intercept parameter estimates were all a little below 0, these tests were slightly more difficult than average. This result is consistent with the result from item P+ values presented above. However, item fit tests indicate that 22 items out of 97 items did not fit well. Appendix C lists the item fit statistics for all items.

Figures 3 and 4 show the distributions of the EAP theta estimates for PAA-A and PAA-B, respectively. One can see that the two theta groups just below 0 had the most students. For PAA-A, the mean and standard deviation of the EAP theta estimates were .05 and 1.01, respectively, and for PAA-B, they were .04 and 1.00, respectively. For both PAAs, their theta estimates had high reliabilities at .90 and .91, respectively. Theta reliability for each test was estimated using the formula below (Haberman & Sinharay, 2010):

$$\hat{R} = 1 - \frac{N^{-1}\sum_{j=1}^{N}\hat{Var}(\theta_j)}{\hat{Var}(\boldsymbol{\theta})},$$

where $\hat{Var}(\theta_j)$ is the estimated posterior variance of examinee j's theta, $\hat{Var}(\boldsymbol{\theta})$ is the estimated posterior population variance of theta, and N is the total number of examinees.

Figures 5 and 6 show the test characteristic curves and test information curves, respectively, for PAA-A and PAA-B based on the EAP theta estimates and EAP true score estimates. For both PAAs, the highest information for PAA-A was a little below 0 and for PAA-B, a little above 0. PAA-B had higher test information than PAA-A across all theta points.

Table 21

PAA-A: Item Parameter Estimates and Standard Errors From Concurrent Calibration

Item	Slope		Intercept 1		Intercept 2		Fit test sig. (chi-square)
	Est.	SE	Est.	SE	Est.	SE	
R_SEASONS11	1.53	.11	1.39	.09	–	–	
R_SEASONS12	1.17	.08	-1.09	.08	–	–	*
R_SEASONS13	.72	.05	-.03	.07	-.40	.08	
R_SEASONS14	.89	.07	.01	.06	–	–	
R_SEASONS15	.86	.07	-1.27	.08	–	–	
R_SEASONS16	.67	.06	-.74	.06	–	–	
R_SEASONS17	.94	.07	-.41	.06	–	–	
R_SEASONS18	.88	.07	-1.15	.07	–	–	*
R_SEASONS19	.68	.04	-.55	.08	-.21	.07	
R_SEASONS21	.82	.05	-.33	.10	.91	.08	*
R_SEASONS22	.35	.04	.04	.06	-1.45	.10	
R_SEASONS23	.58	.05	.55	.07	-.60	.09	**
R_SEASONS24	.45	.05	1.67	.10	1.49	.10	
R_SEASONS25	1.36	.09	-1.12	.08	–	–	
R_SEASONS31	.62	.06	-.38	.06	–	–	
R_SEASONS32	1.59	.09	-.49	.07	–	–	
R_SEASONS41	.25	.03	-1.12	.09	.06	.06	
R_SEASONS42	.23	.04	-.52	.07	-1.00	.08	
R_SEASONS43	.60	.04	-.92	.09	-.34	.07	**
R_SEASONS44	.39	.04	-1.37	.08	-1.24	.08	
CBAL_R_Equating_BA01	.68	.07	.68	.07	–	–	
CBAL_R_Equating_BA02	1.15	.08	.60	.07	–	–	
CBAL_R_Equating_BA03	1.08	.06	.70	.09	.64	.09	
CBAL_R_Equating_BA04	1.16	.08	.55	.07	–	–	
CBAL_R_Equating_BA05	.69	.07	-1.16	.07	–	–	
CBAL_R_Equating_BA06	.52	.06	-.54	.06	–	–	**
CBAL_R_Equating_BA07	1.98	.13	1.31	.10	–	–	
CBAL_R_Equating_BA08	1.25	.09	.53	.07	–	–	
CBAL_R_Equating_BA09	.40	.06	-.63	.06	–	–	*
CBAL_R_Equating_BA10	1.33	.09	-.28	.07	–	–	
CBAL_R_Equating_BA11	.97	.08	-1.22	.08	–	–	**
CBAL_R_Equating_BA12	.52	.10	-2.55	.12	–	–	
CBAL_R_Equating_BA13	1.41	.09	.25	.07	–	–	
CBAL_R_Equating_BA14	.51	.06	-.73	.06	–	–	**
CBAL_R_Equating_BA15	1.36	.09	.10	.07	–	–	
CBAL_R_Equating_BA16	.95	.06	.14	.08	-.71	.09	**
CBAL_R_Equating_BA17	1.28	.08	-.62	.07	–	–	*
CBAL_R_Equating_BA18	1.55	.11	1.42	.09	–	–	
Mean	.90	.07	-.24	.08	-.24	.08	
Standard deviation	.42	.02	.91	.01	.88	.01	
Min	.23	.03	-2.55	.06	-1.45	.06	
Max	1.98	.13	1.67	.12	1.49	.10	

Note. PAA = periodic accountability assessment.

* $p < .05$, ** $p < .01$.

Table 22

PAA-B: Item Parameter Estimates and Standard Errors From Concurrent Calibration

Item	Slope		Intercept 1		Intercept 2		Fit test sig. (chi-square)
	Est.	SE	Est.	SE	Est.	SE	
R_WIND_POWER_11	1.10	.08	-.44	.07	–	–	
R_WIND_POWER_12	1.50	.09	-.20	.07	–	–	
R_WIND_POWER_13	.59	.04	-1.14	.09	-.29	.07	*
R_WIND_POWER_14	.44	.06	-.52	.06	–	–	
R_WIND_POWER_21	.38	.06	-.42	.06	–	–	
R_WIND_POWER_22	1.11	.08	.16	.06	–	–	
R_WIND_POWER_23	.19	.06	-.65	.06	–	–	*
R_WIND_POWER_24	.68	.07	.20	.06	–	–	
R_WIND_POWER_31	1.38	.09	-.14	.07	–	–	
R_WIND_POWER_32	1.15	.08	.33	.07	–	–	
R_WIND_POWER_33	1.22	.07	.31	.08	-.43	.09	
R_WIND_POWER_34	1.26	.07	1.13	.09	.45	.10	
R_WIND_POWER_41	1.00	.06	.72	.08	-.48	.10	
R_WIND_POWER_42	.61	.08	-1.69	.08	–	–	
R_WIND_POWER_43	.70	.05	-.27	.07	-.65	.08	*
R_WIND_POWER_44	1.11	.07	-1.83	.09	-2.47	.14	*
R_WIND_POWER_51	1.08	.09	-1.82	.10	–	–	
R_WIND_POWER_52	1.73	.15	-3.58	.21	–	–	
R_WIND_POWER_53	1.17	.08	-.47	.07	–	–	
R_WIND_POWER_54	1.43	.08	-.63	.08	-1.47	.11	
CBAL_R_Equating_BB01	.53	.06	.00	.06	–	–	
CBAL_R_Equating_BB02	1.77	.11	.21	.08	–	–	
CBAL_R_Equating_BB03	1.01	.08	-.27	.07	–	–	**
CBAL_R_Equating_BB04	1.88	.14	1.98	.12	–	–	
CBAL_R_Equating_BB05	1.10	.08	.32	.07	–	–	
CBAL_R_Equating_BB06	.82	.05	-.93	.09	-.78	.08	*
CBAL_R_Equating_BB07	.82	.05	.50	.09	.63	.09	
CBAL_R_Equating_BB08	.66	.07	.25	.06	–	–	
CBAL_R_Equating_BB09	1.04	.08	.44	.07	–	–	
CBAL_R_Equating_BB10	1.05	.08	-.03	.07	–	–	
CBAL_R_Equating_BB11	1.35	.09	.77	.08	–	–	
CBAL_R_Equating_BB12	1.47	.09	-.45	.07	–	–	
CBAL_R_Equating_BB13	1.24	.09	.77	.07	–	–	
CBAL_R_Equating_BB14	.28	.06	-.86	.07	–	–	
CBAL_R_Equating_BB15	1.13	.08	.46	.07	–	–	
CBAL_R_Equating_BB16	1.08	.08	-.83	.07	–	–	
CBAL_R_Equating_BB17	2.18	.13	-.66	.08	–	–	*
CBAL_R_Equating_BB18	1.32	.10	.87	.08	–	–	*
Mean	1.07	.08	-.22	.08	-.61	.10	
Standard deviation	.44	.02	.98	.03	.94	.02	
Min	.19	.04	-3.58	.06	-2.47	.07	
Max	2.18	.15	1.98	.21	.63	.14	

Note. PAA = periodic accountability assessment.

* $p < .05$, ** $p < .01$.

Table 23

Linking Sets (Blocks C1 and C2): Item Parameter Estimates and Standard Errors From Concurrent Calibration

Item	Slope		Intercept 1		Intercept 2		Fit test sig. (chi-square)
	Est.	SE	Est.	SE	Est.	SE	
CBAL_R_Equating_C01	.91	.07	-.15	.06	–	–	
CBAL_R_Equating_C02	2.21	.15	2.06	.13	–	–	
CBAL_R_Equating_C03	1.44	.09	-.02	.07	–	–	*
CBAL_R_Equating_C04	.85	.05	-2.03	.12	-1.06	.09	
CBAL_R_Equating_C05	1.42	.09	.51	.07	–	–	
CBAL_R_Equating_C06	.45	.06	-.43	.06	–	–	
CBAL_R_Equating_C07	1.06	.09	.86	.07	–	–	
CBAL_R_Equating_C08	.78	.05	-.42	.08	-.59	.08	
CBAL_R_Equating_C10	1.33	.08	1.32	.13	2.23	.13	
CBAL_R_Equating_C11	.78	.07	-1.10	.07	–	–	**
CBAL_R_Equating_C12	.75	.07	-1.02	.07	–	–	
CBAL_R_Equating_C13	.65	.05	.59	.07	-1.02	.10	
CBAL_R_Equating_C14	.42	.04	.48	.07	-.25	.08	
CBAL_R_Equating_C15	.71	.05	-2.63	.13	-1.79	.10	
CBAL_R_Equating_C16	1.37	.08	-.32	.10	.12	.09	*
CBAL_R_Equating_C18	1.55	.10	-.19	.07	–	–	
CBAL_R_Equating_C19	1.22	.07	-.72	.10	-.25	.09	
CBAL_R_Equating_C20	.95	.06	.28	.08	-.36	.09	
CBAL_R_Equating_C21	.82	.05	-1.51	.09	-1.69	.10	
CBAL_R_Equating_C22	.97	.08	-1.24	.08	–	–	
CBAL_R_Equating_C23	1.33	.09	-1.38	.09	–	–	
Mean	1.05	.07	-.34	.09	-.47	.10	
Standard deviation	.42	.02	1.13	.02	1.14	.01	
Min	.42	.04	-2.63	.06	-1.79	.08	
Max	2.21	.15	2.06	.13	2.23	.13	

* $p < .05$, ** $p < .01$.

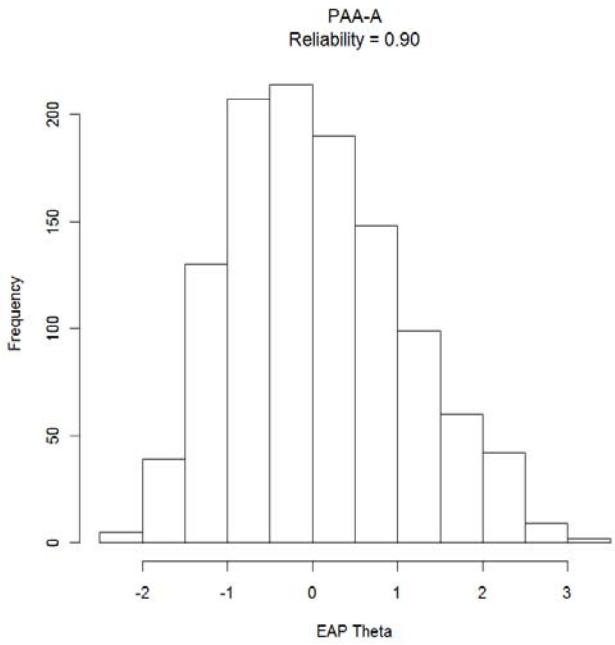

***Figure 3*. PAA-A EAP theta estimate distribution. EAP = expected a posteriori; PAA = periodic accountability assessment.**

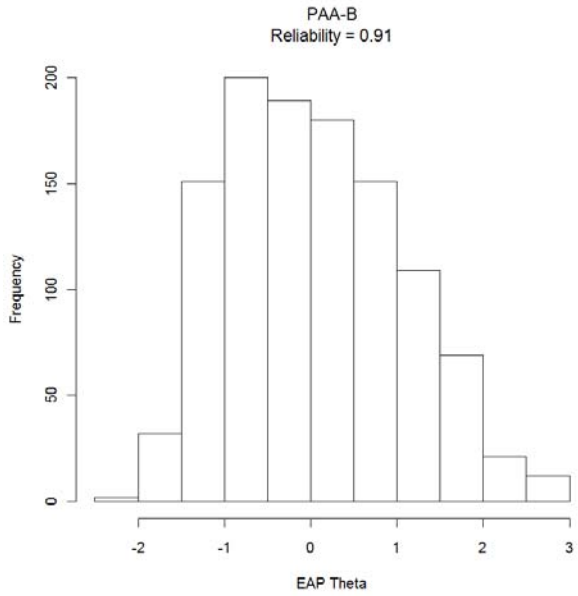

***Figure 4*. PAA-B EAP theta estimate distribution. EAP = expected a posteriori; PAA = periodic accountability assessment.**

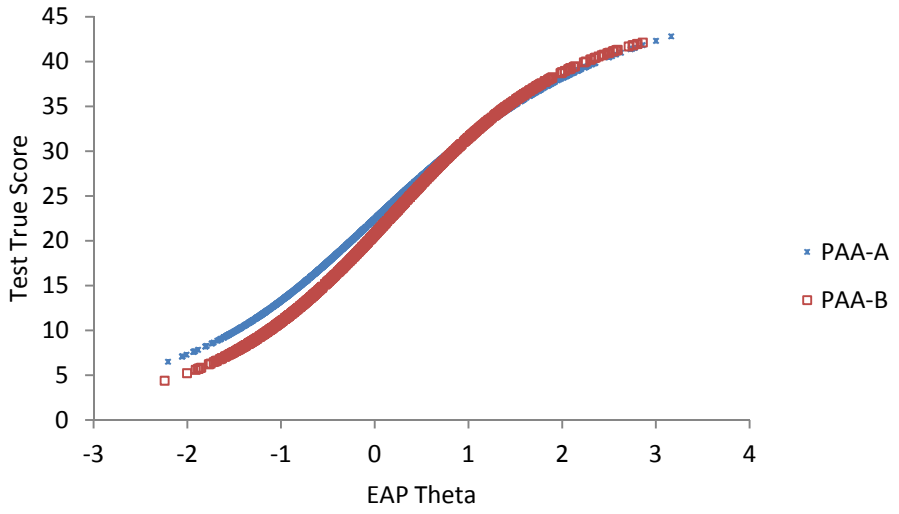

***Figure 5*. Test characteristic curve based on EAP thetas and EAP true scores estimates. EAP = expected a posteriori; PAA = periodic accountability assessment.**

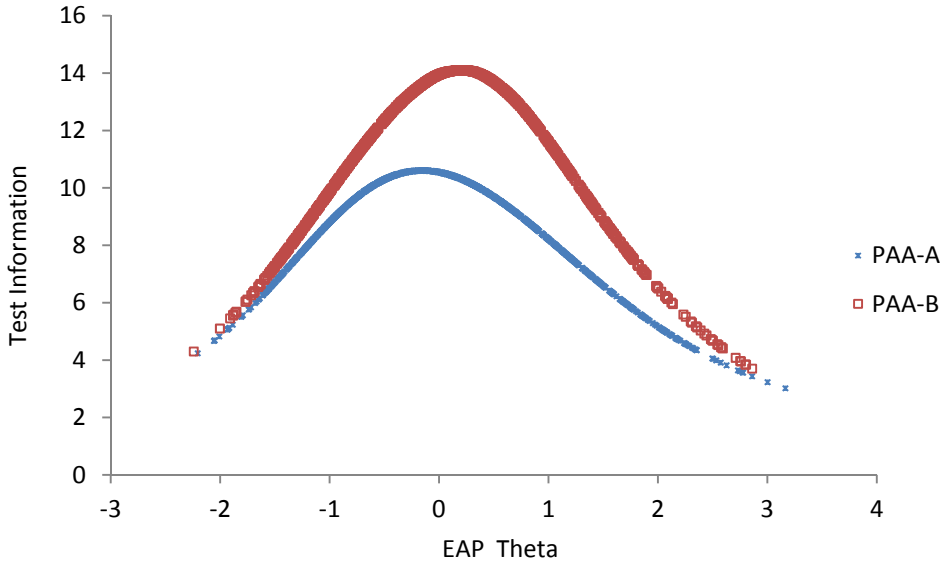

***Figure 6*. Test information curve based on EAP thetas estimates. EAP = expected a posteriori; PAA = periodic accountability assessment.**

Analyses of Factors Affecting Test Scores

The effects of PAA, test-order, and demographic groups on test scores were evaluated using *t*-tests, one way ANOVAs, multiple comparisons, and mixed models.

Subgroup Comparisons

Table 24 provides *t*-test results, as well as means and standard deviations of raw scores and theta estimates on each PAA for several demographic groups: gender, SES, and ELL status. A statistically significant difference was found for SES and ELL groups on both PAAs. The economically disadvantaged and ELL groups had lower mean scores and theta values than their respective counterparts on both PAAs.

Table 24

Subgroup Comparisons on Each PAA

Subgroup	Category	*N*	Theta				Raw score			
			Mean	SD	*t*-value	*p*-value	Mean	SD	*t*-value	*p*-value
PAA-A										
Gender	M	542	.02	1.02	-1.74	.08	22.47	9.84	-1.63	.10
	F	580	.12	.99			23.41	9.45		
Low SES status	Y	468	-.20	.91	7.41	.00**	20.47	8.73	7.20	.00**
	N	344	.34	1.09			25.47	10.46		
ELL status	N	857	.18	1.00	4.10	.00**	23.94	9.57	3.93	.00**
	Y	35	-.53	.88			17.49	8.47		
PAA-B										
Gender	M	544	.05	1.03	.02	.98	21.60	10.79	.22	.83
	F	553	.05	.97			21.46	10.19		
Low SES status	Y	459	-.26	.89	8.93	.00**	18.28	9.39	8.98	.00**
	N	335	.37	1.05			24.90	10.85		
ELL status	N	845	.16	.99	5.60	.00**	22.68	10.36	5.46	.00**
	Y	35	-.54	.71			15.17	7.85		

Note. PAA = periodic accountability assessment.

** $p < .01$.

Because race had four subgroups with unequal variances, Welch's one-way ANOVAs were first run on ethnic groups for theta values and raw scores, respectively, on each PAA. Note that the Native American subgroup was excluded in the one-way ANOVAs and the mixed models below because of small sample size (9 Native American students). As shown in Table 25, all Welch's ANOVA tests were significant. Therefore, multiple comparisons (Dunnett's C test) were conducted on all pairs of racial/ethnic groups. Group pairs having significant differences

are shown in Table 25. Table 25 also provides the mean and standard deviations of the theta estimates and raw scores for each racial/ethnic group in each PAA. One can see that across both PAAs, the order of the test scores of the four ethnic groups from high to low was Asian/Pacific Islander, White, Hispanic, and African American, with most of the score differences between ethnic groups being statistically significant.

Table 25

Race Subgroup Comparisons on Each PAA

Race[a]	*N*	Theta				Raw score				Multiple comparisons[c]		
		Mean	SD	F value[b]	*p*-value	Mean	SD	F value[b]	*p*-value	1	2	3
						PAA-A						
1	37	.94	1.08			31.24	10.29					
2	291	-.32	.78	42.23	.00**	19.23	7.44	41.99	.00**	*		
3	563	.31	1.00			25.22	9.70			*	*	
4	37	-.05	.87			21.46	8.40			*		
						PAA-B						
1	34	.82	1.21			29.56	11.95					
2	273	-.40	.77	51.71	.00**	16.58	8.05	55.93	.00**	*		
3	565	.33	.96			24.60	10.11				*	
4	37	.00	.80			21.03	8.92			*	*	

Note. PAA = periodic accountability assessment.

[a]1 = Asian/Pacific Islander, 2 = African American, 3 = White, 4 = Hispanic. [b] F value from Welch's ANOVA because Levene's test indicated unequal variances among race subgroups. [c] Dunnett's C test; the results from the theta estimates and raw scores were the same.

* $p < .05$, ** $p < .01$.

Mixed Models

Mixed models were used to check the school, PAA, and test-order effects on test scores. The dependent variable was students' theta estimates on each PAA from the GPCM IRT calibrations. In the full model, the random effects were school and student within school, and the fixed effects were PAA (A or B), test-order (Test 1 or Test 2), and their interaction effect. Since the interaction was not significant ($p = .44$), it was dropped from the full model. The model comparisons show that school and student within school were significant random effects (both *p*s = .00). The final model estimates are shown in Table 26, which indicates that test-order was a

significant effect, while PAA was not. Table 27 shows that students performed better on the first PAA than the second PAA, no matter which PAA they took first. We also added the demographic variables to the final model to compare subgroup performance (see Table 28). The test-order was still significant, and SES status, ELL status, and race/ethnicity were also statistically significant. Note that SES status, ELL status, and race/ethnicity were also significant in the *t*-test or one-way ANOVA of population-group means presented earlier.

Table 26

Mixed Model for PAA and Test-Order Effects (N = 2,223)

Fixed effect	Numerator DF	Denominator DF	*F* value	*p*- value	Random effect	Variance
PAA	1	979	1.17	.28	School	.17
Order	1	979	304.44	.00	Student nested in School	.65
					Residual	.16

Note. DF = degree of freedom, PAA = periodic accountability assessment.

Table 27

Mean and Standard Deviation of Theta Estimates by Test Order

Test-order	PAA-A		PAA-B		Total	
	Mean	SD	Mean	SD	Mean	SD
1	.14	.97	.20	.97	.17	.97
2	-.04	1.05	-.16	1.00	-.10	1.03
Total	.05	1.01	.04	1.00	.05	1.01

Note. PAA = periodic accountability assessment.

Table 28

Mixed Model With Subgroup Comparisons (N = 1,130)

Fixed effect	Numerator DF	Denominator DF	F value	*p*- value	Random effect	Variance
PAA	1	525	.64	.43	School	.10
Order	1	525	83.40	.00	Student nested in School	.55
Gender	1	525	2.67	.10	Residual	.15
SES status	1	525	12.95	.00		
ELL status	1	525	7.37	.01		
Race	3	525	29.02	.00		

Note. DF = degree of freedom, ELL = English language learner, PAA = periodic accountability assessment, SES = social economic status.

Summary

The psychometric properties of the fall 2009 CBAL reading PAAs were studied under both classical test theory and item response theory. Classical item statistics and IRT item parameter estimates were reported. Summary statistics and reliabilities of raw subscores and total scores and IRT theta scores were presented. In addition, the report explored the effects of various factors on item and test scores, such as school, PAA, test-order, task, item, student, or demographic groups. The main findings in this report are as follows:

1. At the item level, classical statistics show all items performed reasonably well except for one item, CBAL_R_Equating_C09 in linking Block C1, which had close to 0 correlations with the total scores of PAA-A and PAA-B, and was removed from the test analyses. For the human-scored items, rater reliabilities were very high as indicated by kappa coefficients, rater agreement, and generalizability coefficients. The missing response rates were very low, with all less than or equal to 3.21%. There were only three items identified as having Category-C DIF. The item skill-level classification for PAA-B was reasonable; however, for PAA-A, the items or their classifications could be improved, as the Level 2 items were unexpectedly more difficult than the Level 3 items.

2. At the aggregated score level, the total raw scores of PAA-A and PAA-B had high reliabilities of .89 and .91, respectively. PAA-A and PAA-B had similar difficulty (slightly below average), and their raw score correlation was .79. The intersubscore correlations for subscores with mutually exclusive items ranged between .42 and .75; the intertask correlations in *Seasons* and *Wind Power* were low to moderate in the range between .33 and .64. The similarity between the testlet reliabilities based on task scores and the regular reliabilities based on item scores suggested that task did not appear to have strong testlet effects in PAA-A, PAA-B, *Seasons*, or *Wind Power*.

3. The total raw scores of PAA-A and PAA-B had moderate to high correlations with state ELA tests and reading tests and slightly lower correlations with state math tests, which provides evidence to support the construct validity of both PAAs.

4. At the test level, the test-order, school, student, SES status, ELL status, and race/ethnicity had significant effects on test scores, while test form and gender were

not statistically significant. Students performed better on the first test than the second test, no matter which PAA-order they took. This test-order effect may be due to motivation or fatigue, since there were no stakes associated with performance.

5. Item parameter estimates and EAP theta estimates were very similar between the separate calibration via anchor linking and the concurrent calibration based on the generalized partial credit model. The difference in student performance between the two test occasions seemed to have little effect on the item parameter estimates in the concurrent calibration where students' abilities were assumed equal across the two occasions.
6. IRT results from the concurrent calibration showed all item parameter estimates to be within a reasonable range, and the means of item parameter estimates of PAA-A, PAA-B, and linking blocks to be similar. However, some items did not fit the GPCM well. The reliabilities of the theta estimates for PAA-A and PAA-B were .90 and .91, respectively. PAA-B had higher test information than PAA-A at most of the theta points.
7. Strong correlations and similar distributions of item parameter estimates between the linking sets and the operational forms (PAA-A and PAA-B), as well as no drift of any linking item between the PAA forms, indicate that the linking Blocks C1 and C2 performed well as a linking set.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London, England: Chapman and Hall.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8,* 70–91.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

CBAL ELA Team. (2011). *Reading summative final report 2010*. Unpublished manuscript. Princeton, NJ: ETS.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Deane, P., Fowles, M., & Bennett, R. E. (2009, June). *Cognitively based assessment of, for, and as learning: The case of writing assessment*. Paper presented at the meeting of the National Conference on Student Assessment of the Council for Chief State School Officers, Los Angeles, CA.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209–227.

Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

O'Reilly, T., & Sheehan, K. M. (2009a). *Cognitively based assessment of, for, and as learning: A framework for assessing reading competency* (ETS Research Report No. RR-09–26). Princeton, NJ: ETS.

O'Reilly, T., & Sheehan, K. M. (2009b). *Cognitively based assessment of, for, and as learning: A 21st century approach for assessing reading competency* (ETS Research Memorandum No. 09–04). Princeton, NJ: ETS.

Robelen, E. W. (2010). *Two state groups win federal grants for common tests*. Retrieved from http://www.edweek.org/ew/articles/2010/09/02/03assess.h30.html

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Stocking, M. L., & Lord, F. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Yoo, H., Fu, J., Wise, M. D., & Chung, S. (2011). *Dimensionality analysis of CBAL reading tests*. Manuscript submitted for publication.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Appendix A

Item Score Frequency Tables

Table A1

PAA-A: Item Score Frequency

Item score ID	Score													
	Total		0		1		2		Omit		Not reached		System error	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
R_SEASONS11	1,223	100	328	26.82	890	72.77	–	–	5	.41	–	–	–	–
R_SEASONS12	1,223	100	853	69.75	370	30.25	–	–	–	–	–	–	–	–
R_SEASONS13	1,223	100	476	38.92	391	31.97	356	29.11	–	–	–	–	–	–
R_SEASONS14	1,223	100	602	49.22	617	50.45	–	–	4	.33	–	–	–	–
R_SEASONS15	1,223	100	910	74.41	312	25.51	–	–	–	–	1	.08	–	–
R_SEASONS16	1,223	100	802	65.58	419	34.26	–	–	–	–	2	.16	–	–
R_SEASONS17	1,223	100	710	58.05	511	41.78	–	–	–	–	2	.16	–	–
R_SEASONS18	1,223	100	882	72.12	339	27.72	–	–	–	–	2	.16	–	–
R_SEASONS19	1,223	100	519	42.44	255	20.85	447	36.55	–	–	2	.16	–	–
R_SEASONS21	1,223	100	360	29.44	178	14.55	682	55.76	–	–	3	.25	–	–
R_SEASONS22	1,223	100	532	43.5	548	44.81	140	11.45	–	–	3	.25	–	–
R_SEASONS23	1,223	100	383	31.32	600	49.06	237	19.38	–	–	3	.25	–	–
R_SEASONS24	1,223	100	126	10.3	582	47.59	512	41.86	–	–	3	.25	–	–
R_SEASONS25	1,223	100	843	68.93	376	30.74	–	–	–	–	4	.33	–	–
R_SEASONS31	1,223	100	709	57.97	508	41.54	–	–	–	–	6	.49	–	–
R_SEASONS32	1,223	100	708	57.89	508	41.54	–	–	–	–	7	.57	–	–
R_SEASONS41	1,223	100	504	41.21	161	13.16	546	44.64	–	–	11	.90	1	.08
R_SEASONS42	1,223	100	608	49.71	363	29.68	235	19.22	–	–	17	1.39	–	–
R_SEASONS43	1,223	100	563	46.03	201	16.43	439	35.9	–	–	20	1.64	–	–
R_SEASONS44	1,223	100	755	61.73	192	15.7	256	20.93	–	–	20	1.64	–	–
CBAL_R_Equating_BA02	1,177	100	449	38.15	726	61.68	–	–	–	–	–	–	2	.17
CBAL_R_Equating_BA03	1,177	100	312	26.51	384	32.63	480	40.78	–	–	–	–	1	.08
CBAL_R_Equating_BA04	1,177	100	460	39.08	717	60.92	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA05	1,177	100	866	73.58	311	26.42	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA06	1,177	100	729	61.94	448	38.06	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA07	1,177	100	368	31.27	809	68.73	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA08	1,177	100	469	39.85	708	60.15	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA09	1,177	100	758	64.4	419	35.6	–	–	–	–	–	–	–	–
CBAL_R_Equating_BA10	1,177	100	647	54.97	529	44.94	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA11	1,177	100	855	72.64	320	27.19	–	–	–	–	1	.08	1	.08
CBAL_R_Equating_BA12	1,177	100	1078	91.59	98	8.33	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA13	1,177	100	535	45.45	640	54.38	–	–	–	–	1	.08	1	.08
CBAL_R_Equating_BA14	1,177	100	776	65.93	400	33.98	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA15	1,177	100	566	48.09	610	51.83	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA16	1,177	100	463	39.34	424	36.02	289	24.55	–	–	1	.08	–	–
CBAL_R_Equating_BA17	1,177	100	718	61	458	38.91	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA18	1,177	100	314	26.68	862	73.24	–	–	–	–	1	.08	–	–
CBAL_R_Equating_BA01	1,177	100	406	34.49	769	65.34	–	–	–	–	2	.17	–	–

Note. PAA = periodic accountability assessment.

Table A2

PAA-B: Item Score Frequency

Item score ID	Total		Score											
			0		1		2		Omit		Not reached		System error	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
R_WIND_POWER_11	1,214	100	710	58.48	504	41.52	–	–	–	–	–	–	–	–
R_WIND_POWER_12	1,214	100	652	53.71	562	46.29	–	–	–	–	–	–	–	–
R_WIND_POWER_13	1,214	100	582	47.94	166	13.67	466	38.39	–	–	–	–	–	–
R_WIND_POWER_14	1,214	100	750	61.78	464	38.22	–	–	–	–	–	–	–	–
R_WIND_POWER_21	1,214	100	726	59.8	488	40.2	–	–	–	–	–	–	–	–
R_WIND_POWER_22	1,214	100	567	46.71	647	53.29	–	–	–	–	–	–	–	–
R_WIND_POWER_23	1,214	100	795	65.49	419	34.51	–	–	–	–	–	–	–	–
R_WIND_POWER_24	1,214	100	549	45.22	665	54.78	–	–	–	–	–	–	–	–
R_WIND_POWER_31	1,214	100	634	52.22	576	47.45	–	–	3	.25	1	.08	–	–
R_WIND_POWER_32	1,214	100	526	43.33	687	56.59	–	–	–	–	1	.08	–	–
R_WIND_POWER_33	1,214	100	446	36.74	413	34.02	348	28.67	5	.41	2	.16	–	–
R_WIND_POWER_34	1,214	100	298	24.55	514	42.34	399	32.87	–	–	3	.25	–	–
R_WIND_POWER_41	1,214	100	372	30.64	566	46.62	272	22.41	–	–	4	.33	–	–
R_WIND_POWER_42	1,214	100	994	81.88	213	17.55	–	–	–	–	6	.49	1	.08
R_WIND_POWER_43	1,214	100	526	43.33	356	29.32	325	26.77	1	.08	6	.49	–	–
R_WIND_POWER_44	1,214	100	852	70.18	146	12.03	204	16.8	3	.25	9	.74	–	–
R_WIND_POWER_51	1,214	100	969	79.82	229	18.86	–	–	–	–	16	1.32	–	–
R_WIND_POWER_52	1,214	100	1089	89.7	103	8.48	–	–	–	–	22	1.81	–	–
R_WIND_POWER_53	1,214	100	698	57.5	491	40.44	–	–	–	–	25	2.06	–	–
R_WIND_POWER_54	1,214	100	620	51.07	260	21.42	295	24.3	9	.74	30	2.47	–	–
CBAL_R_Equating_BB02	1,153	100	540	46.83	613	53.17	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB03	1,153	100	634	54.99	519	45.01	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB04	1,153	100	249	21.6	904	78.4	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB05	1,153	100	500	43.37	653	56.63	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB06	1,153	100	598	51.86	205	17.78	350	30.36	–	–	–	–	–	–
CBAL_R_Equating_BB07	1,153	100	301	26.11	362	31.4	490	42.5	–	–	–	–	–	–
CBAL_R_Equating_BB08	1,153	100	506	43.89	647	56.11	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB09	1,153	100	470	40.76	683	59.24	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB10	1,153	100	579	50.22	574	49.78	–	–	–	–	–	–	–	–
CBAL_R_Equating_BB11	1,153	100	414	35.91	738	64.01	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB12	1,153	100	665	57.68	487	42.24	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB13	1,153	100	408	35.39	744	64.53	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB14	1,153	100	801	69.47	350	30.36	–	–	–	–	1	.09	1	.09
CBAL_R_Equating_BB15	1,153	100	468	40.59	684	59.32	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB16	1,153	100	756	65.57	396	34.35	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB17	1,153	100	686	59.5	466	40.42	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB18	1,153	100	393	34.08	759	65.83	–	–	–	–	1	.09	–	–
CBAL_R_Equating_BB01	1,153	100	570	49.44	582	50.48	–	–	–	–	1	.09	–	–

Note. PAA = periodic accountability assessment.

Table A3

Linking Blocks C1 and C2: Item Score Frequency

Section	Item score ID	Score											
		Total		0		1		2		Not reached		System error	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
	CBAL_R_Equating_C01	1,172	100	617	52.65	554	47.27	–	–	1	.09	–	–
	CBAL_R_Equating_C05	1,172	100	479	40.87	692	59.04	–	–	1	.09	–	–
	CBAL_R_Equating_C06	1,172	100	699	59.64	472	40.27	–	–	1	.09	–	–
	CBAL_R_Equating_C08	1,172	100	528	45.05	296	25.26	347	29.61	1	.09	–	–
	CBAL_R_Equating_C09	1,172	100	981	83.7	190	16.21	–	–	1	.09	–	–
C1	CBAL_R_Equating_C10	1,172	100	191	16.3	281	23.98	699	59.64	1	.09	–	–
	CBAL_R_Equating_C13	1,172	100	382	32.59	622	53.07	167	14.25	1	.09	–	–
	CBAL_R_Equating_C15	1,172	100	867	73.98	64	5.46	240	20.48	1	.09	–	–
	CBAL_R_Equating_C16	1,172	100	479	40.87	199	16.98	493	42.06	1	.09	–	–
	CBAL_R_Equating_C22	1,172	100	857	73.12	313	26.71	–	–	1	.09	1	.09
	CBAL_R_Equating_C23	1,172	100	857	73.12	314	26.79	–	–	1	.09	–	–
	CBAL_R_Equating_C02	1,158	100	268	23.14	890	76.86	–	–	–	–	–	–
	CBAL_R_Equating_C03	1,158	100	582	50.26	576	49.74	–	–	–	–	–	–
	CBAL_R_Equating_C04	1,158	100	714	61.66	85	7.34	358	30.92	–	–	1	.09
	CBAL_R_Equating_C07	1,158	100	380	32.82	778	67.18	–	–	–	–	–	–
	CBAL_R_Equating_C11	1,158	100	832	71.85	326	28.15	–	–	–	–	–	–
C2	CBAL_R_Equating_C12	1,158	100	818	70.64	340	29.36	–	–	–	–	–	–
	CBAL_R_Equating_C14	1,158	100	345	29.79	528	45.6	284	24.53	–	–	1	.09
	CBAL_R_Equating_C18	1,158	100	616	53.2	542	46.8	–	–	–	–	–	–
	CBAL_R_Equating_C19	1,158	100	529	45.68	174	15.03	455	39.29	–	–	–	–
	CBAL_R_Equating_C20	1,158	100	413	35.66	414	35.75	330	28.5	–	–	1	.09
	CBAL_R_Equating_C21	1,158	100	754	65.11	165	14.25	239	20.64	–	–	–	–

Appendix B

Item DIF Results

Table B1

PAA-A: Item DIF Categories

Item score ID	Male (N = 542) vs. Female (N = 580)	White (N = 563) vs. Black (N = 291)	White (N = 563) vs. Combination[a] (N = 83)	Low SES status: No (N = 344) vs. Yes (N = 468)	Number of C DIF (if not 0)
R_SEASONS11	A	A	A	A	
R_SEASONS12	A	A	A	A	
R_SEASONS13	A	A	A	A	
R_SEASONS14	A	A	A	A	
R_SEASONS15	A	A	A	A	
R_SEASONS16	A	A	B+	A	
R_SEASONS17	A	A	B-	A	
R_SEASONS18	A	A	A	A	
R_SEASONS19	A	A	A	A	
R_SEASONS21	A	A	A	A	
R_SEASONS22	A	A	A	A	
R_SEASONS23	A	B+	A	A	
R_SEASONS24	A	A	A	A	
R_SEASONS25	A	A	A	A	
R_SEASONS31	A	A	A	A	
R_SEASONS32	A	A	A	A	
R_SEASONS41	A	A	A	A	
R_SEASONS42	A	A	A	A	
R_SEASONS43	A	A	A	A	
R_SEASONS44	A	A	A	A	
CBAL_R_Equating_BA01	A	A	A	A	
CBAL_R_Equating_BA02	A	A	A	A	
CBAL_R_Equating_BA03	A	A	A	A	
CBAL_R_Equating_BA04	A	B-	A	A	
CBAL_R_Equating_BA05	A	A	A	A	
CBAL_R_Equating_BA06	A	A	A	A	
CBAL_R_Equating_BA07	A	A	A	A	
CBAL_R_Equating_BA08	A	A	A	A	
CBAL_R_Equating_BA09	B-	A	A	A	
CBAL_R_Equating_BA10	A	A	A	A	
CBAL_R_Equating_BA11	A	A	A	A	
CBAL_R_Equating_BA12	A	A	A	A	
CBAL_R_Equating_BA13	A	A	A	A	
CBAL_R_Equating_BA14	A	A	A	A	
CBAL_R_Equating_BA15	A	A	A	A	
CBAL_R_Equating_BA16	A	A	A	A	
CBAL_R_Equating_BA17	A	A	A	A	
CBAL_R_Equating_BA18	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. PAA = periodic accountability

[a]Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B2

PAA-B: Item DIF Categories

Item score ID	Male (N = 544) vs. Female (N = 553)	White (N = 565) vs. Black (N = 273)	White (N = 565) vs. Combination[a] (N = 80)	Low SES status: No (N = 335) vs. Yes (N = 459)	Number of C DIF (if not 0)
R_WIND_POWER_11	A	A	A	A	
R_WIND_POWER_12	C-	B-	A	A	1
R_WIND_POWER_13	A	B-	A	A	
R_WIND_POWER_14	A	A	A	A	
R_WIND_POWER_21	A	A	B+	A	
R_WIND_POWER_22	A	A	A	A	
R_WIND_POWER_23	A	A	A	A	
R_WIND_POWER_24	B-	A	A	A	
R_WIND_POWER_31	B+	B+	A	A	
R_WIND_POWER_32	A	B+	A	A	
R_WIND_POWER_33	A	A	A	A	
R_WIND_POWER_34	A	A	A	A	
R_WIND_POWER_41	A	A	A	A	
R_WIND_POWER_42	A	A	A	A	
R_WIND_POWER_43	A	A	A	A	
R_WIND_POWER_44	A	A	A	A	
R_WIND_POWER_51	B+	A	A	A	
R_WIND_POWER_52	A	A	A	A	
R_WIND_POWER_53	A	A	A	A	
R_WIND_POWER_54	A	A	A	A	
CBAL_R_Equating_BB01	A	A	B-	A	
CBAL_R_Equating_BB02	A	A	A	A	
CBAL_R_Equating_BB03	A	A	A	A	
CBAL_R_Equating_BB04	A	A	A	B-	
CBAL_R_Equating_BB05	A	B+	A	A	
CBAL_R_Equating_BB06	A	A	A	A	
CBAL_R_Equating_BB07	A	B+	A	A	
CBAL_R_Equating_BB08	A	A	A	A	
CBAL_R_Equating_BB09	A	A	A	A	
CBAL_R_Equating_BB10	B-	A	A	A	
CBAL_R_Equating_BB11	A	A	A	A	
CBAL_R_Equating_BB12	A	A	B+	A	
CBAL_R_Equating_BB13	A	A	A	A	
CBAL_R_Equating_BB14	A	A	A	A	
CBAL_R_Equating_BB15	A	A	A	A	
CBAL_R_Equating_BB16	A	A	A	A	
CBAL_R_Equating_BB17	A	A	A	A	
CBAL_R_Equating_BB18	A	B-	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. PAA = periodic accountability assessment.

[a]Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B3

Linking Block C1: Item DIF Categories

Item score ID	PAA-A: Male (*N* = 516) vs. Female (*N* = 552)	PAA-B: Male (*N* = 514) vs. Female (*N* = 530)	PAA-A: White (*N* = 533) vs. Black (*N* = 276)	PAA-B: White (*N* = 533) vs. Black (*N* = 262)	PAA-A: White (*N* = 533) vs. Combination[a] (*N* = 80)	PAA-B: White (*N* = 533) vs. Combination[a] (*N* = 77)	PAA-A: low SES status: No (*N* = 339) vs. Yes (*N* = 454)	PAA-B: low SES status: No (*N* = 332) vs. Yes (*N* = 449)	Number of C DIF (if not 0)
CBAL_R_Equating_C01	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C05	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C06	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C08	B-	C-	B-	A	C-	C-	A	A	3
CBAL_R_Equating_C10	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C13	A	A	A	A	A	A	A	A	1
CBAL_R_Equating_C15	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C16	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C22	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C23	A	A	A	B+	A	A	A	A	

Note. The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. PAA = periodic accountability assessment.

[a]Combination of Native American, Asian/Pacific Islander, and Hispanic.

Table B4

Linking Block C2: Item DIF Categories [a]

Item score ID	PAA-A: Male (*N* = 515) vs. Female (*N* = 540)	PAA-B: Male (*N* = 516) vs. Female (*N* = 529)	PAA-A: White (*N* = 533) vs. Black (*N* = 270)	PAA-B: White (*N* = 532) vs. Black (*N* = 260)	PAA-A: White (*N* = 533) vs. Combination[b] (*N* = 79)	PAA-B: White (*N* = 532) vs. Combination[b] (*N* = 77)	PAA-A: low SES status: No (*N* = 335) vs. Yes (*N* = 448)	PAA-B: low SES status: No (*N* = 332) vs. Yes (*N* = 439)	Number of C DIF (if not 0)
CBAL_R_Equating_C02	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C03	A	B+	A	A	B+	C+	A	A	1
CBAL_R_Equating_C04	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C07	A	A	B-	A	A	A	A	A	
CBAL_R_Equating_C11	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C12	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C14	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C18	A	A	A	A	B-	A	B-	A	
CBAL_R_Equating_C19	A	A	A	A	A	A	A	A	
CBAL_R_Equating_C20	A	B+	A	A	A	A	A	A	2
CBAL_R_Equating_C21	A	A	A	A	A	A	A	A	

Note. PAA = periodic accountability assessment.

[a]The first group is the reference group, and the second group is the focus group. A positive sign favors the focus group, while a negative sign favors the reference group. [b]Combination of Native American, Asian/Pacific Islander, and Hispanic.

Appendix C

IRT Item Fit Statistics

Table C1

PAA-A: Item Fit Statistics

Item	Chi-square	DF	Prob.	Sig.
R_SEASONS11	22.22	19	.27	
R_SEASONS12	47.55	32	.04	*
R_SEASONS13	70.42	54	.07	
R_SEASONS14	37.48	28	.11	
R_SEASONS15	41.00	34	.19	
R_SEASONS16	30.47	34	.64	
R_SEASONS17	24.22	32	.84	
R_SEASONS18	48.92	34	.05	*
R_SEASONS19	56.94	51	.26	
R_SEASONS21	55.90	39	.04	*
R_SEASONS22	49.89	66	.93	
R_SEASONS23	149.12	57	.00	**
R_SEASONS24	52.15	42	.14	
R_SEASONS25	39.63	32	.17	
R_SEASONS31	46.74	33	.06	
R_SEASONS32	18.51	26	.86	
R_SEASONS41	26.68	44	.98	
R_SEASONS42	79.52	66	.12	
R_SEASONS43	83.79	49	.00	**
R_SEASONS44	6.40	52	.20	
CBAL_R_Equating_BA01	27.38	24	.29	
CBAL_R_Equating_BA02	29.25	22	.14	
CBAL_R_Equating_BA03	58.21	45	.09	
CBAL_R_Equating_BA04	22.00	24	.58	
CBAL_R_Equating_BA05	38.07	35	.33	
CBAL_R_Equating_BA06	56.69	33	.01	**
CBAL_R_Equating_BA07	24.32	20	.23	
CBAL_R_Equating_BA08	21.69	23	.54	
CBAL_R_Equating_BA09	54.24	34	.02	*
CBAL_R_Equating_BA10	22.08	27	.73	
CBAL_R_Equating_BA11	74.52	34	.00	**
CBAL_R_Equating_BA12	40.62	37	.31	
CBAL_R_Equating_BA13	29.05	24	.22	
CBAL_R_Equating_BA14	70.49	35	.00	**
CBAL_R_Equating_BA15	30.09	24	.18	
CBAL_R_Equating_BA16	107.90	54	.00	**
CBAL_R_Equating_BA17	45.11	29	.03	*
CBAL_R_Equating_BA18	14.58	18	.69	

Note. DF = degree of freedom, PAA = periodic accountability assessment, Prob. = probability, Sig. = significance.

* $p < .05$, ** $p < .01$.

Table C2

PAA-B: Item Fit Statistics

Item	Chi-square	DF	Prob.	Sig.
R_WIND_POWER_11	26.87	29	.58	
R_WIND_POWER_12	18.94	26	.84	
R_WIND_POWER_13	68.02	45	.02	*
R_WIND_POWER_14	31.66	35	.63	
R_WIND_POWER_21	28.23	35	.79	
R_WIND_POWER_22	24.26	28	.67	
R_WIND_POWER_23	51.34	36	.05	*
R_WIND_POWER_24	30.30	30	.45	
R_WIND_POWER_31	37.63	28	.11	
R_WIND_POWER_32	31.33	26	.22	
R_WIND_POWER_33	60.82	50	.14	
R_WIND_POWER_34	51.38	46	.27	
R_WIND_POWER_41	48.19	51	.59	
R_WIND_POWER_42	34.60	36	.54	
R_WIND_POWER_43	85.58	58	.01	*
R_WIND_POWER_44	58.95	42	.04	*
R_WIND_POWER_51	37.89	35	.34	
R_WIND_POWER_52	41.29	36	.25	
R_WIND_POWER_53	34.22	29	.23	
R_WIND_POWER_54	64.79	48	.05	
CBAL_R_Equating_BB01	38.71	31	.16	
CBAL_R_Equating_BB02	24.48	22	.32	
CBAL_R_Equating_BB03	50.86	28	.01	**
CBAL_R_Equating_BB04	10.84	15	.76	
CBAL_R_Equating_BB05	26.61	24	.32	
CBAL_R_Equating_BB06	73.52	48	.01	*
CBAL_R_Equating_BB07	47.12	45	.39	
CBAL_R_Equating_BB08	27.33	27	.45	
CBAL_R_Equating_BB09	21.58	24	.61	
CBAL_R_Equating_BB10	25.13	27	.57	
CBAL_R_Equating_BB11	21.23	20	.38	
CBAL_R_Equating_BB12	38.87	27	.07	
CBAL_R_Equating_BB13	24.94	20	.20	
CBAL_R_Equating_BB14	22.85	35	.94	
CBAL_R_Equating_BB15	27.00	22	.21	
CBAL_R_Equating_BB16	25.96	31	.72	
CBAL_R_Equating_BB17	39.72	25	.03	*
CBAL_R_Equating_BB18	31.91	20	.04	*

Note. DF = degree of freedom, PAA = periodic accountability assessment, Prob. = probability, Sig. = significance.

* $p < .05$, ** $p < .01$.

Table C3

Linking Blocks C1 and C2: Item Fit Statistics

Item	Chi-square	DF	Prob.	Sig.
CBAL_R_Equating_C01	37.23	28	.11	
CBAL_R_Equating_C02	19.41	15	.20	
CBAL_R_Equating_C03	39.09	25	.04	*
CBAL_R_Equating_C04	32.27	34	.55	
CBAL_R_Equating_C05	13.34	21	.90	
CBAL_R_Equating_C06	34.52	33	.40	
CBAL_R_Equating_C07	23.07	20	.29	
CBAL_R_Equating_C08	52.75	51	.41	
CBAL_R_Equating_C10	28.43	32	.65	
CBAL_R_Equating_C11	56.83	34	.01	**
CBAL_R_Equating_C12	40.64	34	.20	
CBAL_R_Equating_C13	66.77	55	.13	
CBAL_R_Equating_C14	72.61	58	.09	
CBAL_R_Equating_C15	42.86	32	.10	
CBAL_R_Equating_C16	57.65	39	.03	*
CBAL_R_Equating_C18	33.60	25	.12	
CBAL_R_Equating_C19	44.90	39	.24	
CBAL_R_Equating_C20	49.66	50	.49	
CBAL_R_Equating_C21	58.98	44	.07	
CBAL_R_Equating_C22	38.81	32	.19	
CBAL_R_Equating_C23	32.61	31	.39	

Note. DF = degree of freedom, Prob. = probability, Sig. = significance.

* $p < .05$, ** $p < .01$.